Compression of DNA Sequence Using Deep LSTM Neural Network

Prashanth .M .C[1]

Department of P.G. Studies and Research in Computer Science,
Kuvempu University, Shankarghatta - 577 451,
Shimoga, Karnataka, INDIA
prashanth.m.c87@gmail.com

Ravikumar M[2]

Department of P.G. Studies and Research in Computer Science,
Kuvempu University, Shankarghatta - 577 451,
Shimoga, Karnataka, INDIA
ravikumar2142@yahoo.co.in

Abstract-

Compression of DNA sequence is rapidly evolving as a field of research. The researchers are persistently analysing the DNA sequences for several purposes. Hence, the DNA sequences have to be stored and transmitted for analysing the DNA sequences. However, the huge size of the DNA sequences leads to a high cost for transmission. Thus, compressing the DNA sequence data is essential to reduce the size, minimise transmission costand help in achieving efficient analysis. This study aimed to set the encoder and decoder for DNA sequence compression. This study proposed a deep LSTM Neural Network to compress the DNA sequence that leads to various merits effectively. Initially, the DNA sequence dataset is taken as input. The data is loaded, and the vectorizer technique is performed using the encoder, channel and decoder. Then, the encoder and decoder are set for compression. The compression process is performed through the proposed Deep LSTM Neural Network. The Deep LSTM Neural Network consists of the LSTM layer, dense layer and hidden layer. The trained model is taken into account to find the compression results. Thus the results are finally analysed to validate its efficiency.

Keywords: DNA sequence compression, Deep LSTM Neural Network, Encoder and Decoder, Minimised Size and Efficient Analysis.

I. INTRODUCTION

Deoxyribonucleic Acid, abbreviated as DNA, comprises of four-bases. They are Adenine (A), Guanine (G), Thymine (T) as well as Cytosine (C). A huge count of DNA sequences requires high storage space. Hence, there arises a need to determine the way to minimise the size. Indeed, data compression is a technique where the data encoding is done in a small bit count, which is utilised by the decrypted data. Recently, the widespread utilisation of the DNA sequences in diverse fields led to rapid growth in the count of such sequences residing in the DNA databases. These sequences have been stored in an uncompressed form. This requires more storage space. Thus, an effective methodology to store huge data in the minimum count of nucleotides have been introduced in this study [1] that utilised Huffman coding of minimum variance. The algorithm affords a high code rate corresponding to the DNA sequences. Thus, this minimised the synthesising cost. It includes compression as well as the variable length of constraint coding. This methodology can be extended further to other multi-media data. Besides, error detection, as well as correction, can be executed in the DNA sequences for making them robust. Moreover, this paper [2] explored that LZW is a universal methodology to compress the textual data. This is used in the compression of huge DNA sequences. Its utility in DNA sequence compression is expected to have huge merits to the research community [3]. Similarly, this paper [4] introduced an ETBWT-II-LCP methodology to assist in compressing as well as indexing huge DNA sequence collections. Following this methodology, decoding time, as well as memory space, is reduced. This proves the efficiency of the proposed method. Likewise, this study explored a DNA compressor – DNAC-SBE. The proposed compressor has been assessed to validate its efficacy. The outcomes revealed that the proposed compressor performed effectively than the traditional compressors with respect to the compression ratio in spite of the data size or file format. The proposed compressor is highly suitable when the storage space, data transfer and compressed data are vital. More attempts are yet to be implemented for optimising the compression speed and memory utilisation [5]. To overcome these issues, this study introduced a Deep LSTM Neural Network to accomplish DNA Sequence compression effectively.

The major contributions of this study are listed below.

- To set the encoder and decoder for compressing the DNA sequences.
- To effectively compress the DNA sequence using the proposed Deep LSTM Neural Network.
- To analyse the compression outcomes of the proposed methodology with respect to time, ratio and loss so as to validate the efficiency of the system.

1.1 Paper Organisation

Section I explores the basic ideas involved in DNA sequence compression. Subsequently, section II explores various reviews of the existing works related to this context. Then, the proposed methodology is explained in section III. The results obtained through the implementation of the proposed system is discussed in section IV. Finally, the overall summary of the proposed system is concluded in section V.

II. REVIEW OF EXISTING WORK

This section reviews the various existing methodologies related to DNA sequence compression.

The rapid development of effective data compressors corresponding to the DNA sequences is vital to minimise the storage and transmission bandwidth. It is mainly needed for analysis. This study [6] recommended a compression algorithm on the basis of DNA sequence binary representation. Initially, a new methodology has been used for DNA sequence compression. Then, it is converted into binary form. The resulting DNA has been compressed by the use of Extended ASCII encoding. Here one character denotes four nucleotides. The proposed methodology is found to be effective in terms of the compression ratio. Yet, this study has to be further improved by employing the proposed method to compression algorithms on the basis of Markov models for prognosticating the frequency of individual nucleotide relying on the genome region. This will help in accomplishing an effective compression rate than the existing methods. Moreover, various compression algorithms intend to compress DNA sequences. This paper [7] found that techniques that rely on multi-reference can afford efficient performance than the existing techniques. Yet, it is ineffective in terms of time complexity. Thus, a better algorithm has to be developed with respect to time complexity and compression rate. In addition, this article [8] explored lossless compressor with enhanced compression abilities for the DNA sequences denoting varied domains. The outcomes revealed that the recommended technique accomplished a better compression ratio when compared to traditional methods using a diverse benchmark. The computational properties required by the introduced strategy are competitive. Similarly, this paper [9] explored that various executions of Huffman encoding include the features of DNA sequence and it is confirmed to exhibit better compression of DNA data. Moreover, this article [10] presented a DNA sequence corpus to perform a compression benchmark. An analysis on the basis of compression has been implemented on the corpus. Thus, the proposed corpus has been used to execute a compression benchmark that exhibits a clear partitioning from XM and GeCo (DNA compressors) about the compression capability. This study has to perform a full comparison of the proposed and existing methods. Accordingly, this paper [11] recommended a Compress Best for DNA sequence compression that aids in accomplishing an effective compression ratio. It is effective than the traditional compression methodologies. Likewise, this paper [12] introduced a DNA sequence compression algorithm. This is found to exhibit an effective compression ratio than the existing methods. However, the decompressing time makes it unfit for huge data. Additionally, this study [13] proposed effective compression of DNA sequences with NNs (Neural Networks). It has been found that the recommended technique is portable and need only the model possibilities as inputs, thereby affording effective adaptation to additional data analysis tools based on compression or data compressors.

For the alphabets in DNA (A – Adenine, C – Cytosine, G – Guanine and T – Thymine), the average length of description corresponding to two bits per base indicates the maximum length needed for DNA encoding. The merits, as well as the demerits of the traditional compression methodologies, have been re-examined, and a new attempt has been initiated. Based on the comparative analysis of the traditional methods [14], a new technique has been introduced to compress the DNA sequence without relying on the sequence set statistics. The existing techniques considered only self-references. Moreover, this study accomplished maximum compression ratio and will be advantageous when cross similarities have been taken into account. In addition, this paper [15] recommended a compression algorithm for DNA sequence compression. This technique considered intrachromosomal and interchromosomal similarities. The results revealed that the extrachromosomal matches are found to be maximum when compared to inter-chromosomal matches. This aids to accomplish an effective compression ratio. On the other hand, the DNA sequences can be completely compressed only when the mature dictionary is organised for replacement purpose. Similarly, this paper [16] presented a representation technique for DNA isolate compression. Effective outcomes have been attained in compressing the DNA sequence. Yet, this

study has to enhance the compression accuracy. Likewise, this article [17] recommended a compression technique termed DeepDNA, which is a hybrid convolutional and recurrent DNN (Deep Neural Network) to accomplish compression of DNA sequence. The experimental outcomes on the DNA datasets exhibit the efficiency of the proposed methodology. This study could show strategies to operate with the lossless compression of DNA sequences in the near future. Moreover, this study [18] examined several methodologies for compressing DNA data to accomplish efficient storage. The analysis explores that the efficient analysis for compressing DNA sequence remains a challenge for the investigators. Accordingly, this article [19] recommended a lossless compression strategy termed CIGAR Coil consists of the DNA reads. The proposed system is analysed to validate its efficiency. The results revealed that the proposed system showed better results than the existing methods. Moreover, this paper [20] introduced a methodology to minimise the DNA size that utilised the RLIBC (Run Length Index-Based Coding) methodology that performs DNA sample encryption, thereby minimises its size. A comparative analysis has been performed by comparing the proposed and existing methods. The analytical results revealed that the proposed system is effective than the existing system with respect to less complexity and effective compression ratio.

Similarly, this paper [21]introduced and devised techniques for compressing the DNA sequence using NN predictors. Efficient results have been obtained through the proposed methodology. Yet, the compressor performance has to be improved so as to enhance the system efficacy. Moreover, this study [22] recommended a lossless compression methodology that compresses data on the basis of two tiers. This proposed selection encryption corresponding to the modified RSA methodology is a strategy for computational resource minimisation for huge sized DNA. The drawback of this study is the compression of DNA sequence is one-pass. On the other hand, this study [23] proposed the selective encryption and repeat methodology for compressing the DNA sequences through the modified Huffman methodology. Thus, effective outcomes have been attained through the proposed method. Likewise, this study [24] recommended an effective and quick sequence compression methodology on the basis of the variable LUT (Look Up Table) and Differential Direct Coding. The outcomes revealed that the recommended algorithm is effective than the traditional method. The suggested algorithm may afford effective compression as the lengthy sequences have been frequently found in the huge sequences. Similarly, this paper [25] recommended a substitution technique for compressing the DNA sequence. This recommended technique has accomplished an effective compression rate, thereby possessing added merits in dynamic programming. Yet, the compression rate has to be upgraded further so as to improve the system efficiency. Additionally, this article [26] explored the dictionary-based methodologies to compress the DNA sequences via different repetitive structures. These are inbuilt within these sequences. Effective outcomes have been attained through the proposed method. Likewise, this study [27] aimed to exhibit the efficiency in compression on a 2L (Two-level) inverted index personalised for the n-gram indices. This study surveyed the inverted index compression by optimising the PFD compression methodology termed Optimal PForDelta (Optimal PForDelta) and Inter-Polative Coding (IPC). The proposed method is analysed to assess the performance efficiency. The results revealed that the proposed system is effective than the existing methods. Yet, this system has to be extended further so as to enhance the compression rate. Accordingly, this paper [28] explored various technologies that intend to minimise the transmission and storage costs and improve the transmission speed and facilitate in dealing with these DNA data by affording an efficient and flexible solution for compression. Yet, this study needs further enhancement in data compression as well as representation. On the contrary, this article [29] recommended an iCGR (integer Chaos Game Representation) of DNA sequences for compression. This proposed methodology affords a potential tool to accomplish sequential analysis as well as operations. On the other hand, this study [30] introduced a NAF (Nucleotide Archival Format) that permits effective lossless and reference-free DNA sequence compression. This also affords an effective integration of compactness as well as effective compression speed that make it best for the data transfer in one time.

III. PROPOSED METHODOLOGY

According to biological researchers, the DNA sequence is the long string comprised of certain significant genetic information. The DNA researchers are recently increasing, and thus the users facing certain challenges in transfer, data storage and maintenance. There is a space requirement for storage because of the greater sequence size. Therefore, a method is required to decrease the amount of space needed. Data compression is an effective process to reduce DNA sequence size. It can result in reduced storage space and the requirement of bandwidth transfer. In this study, the large set of DNA sequences are focused, and the data compression has been performed to reduce the storage space using the Deep LSTM Neural Network, and the flow is shown in figure.1. The data compression result of the DNA sequence is further obtained with respect to compressed ratio, time and loss.

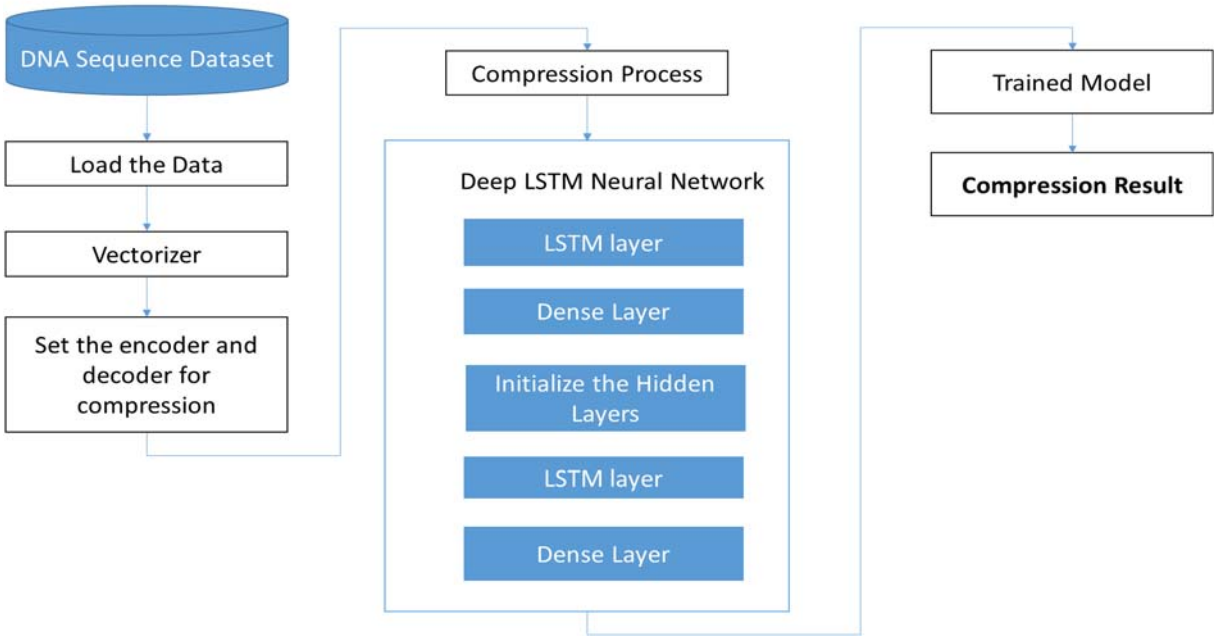

Figure.1. Proposed Flow

In this proposed study, the DNA sequences are input sequences in which the data is loaded initially. Further, the vectorizer technique is performed using the encoder, channel and decoder shown in Figure.2. Encoder comprised with vectors, indexing and codebook generation. The data is divided into non-overlapping blocks. Effective codebook generation is considered a significant task for Vectorization. To the receiver, the index number is transferred through the channel, finally, in a decoder comprised with index table, reconstructed output and codebook. Both transmitter and receiver received a similar codebook. Thus the input data is identical to reconstructed output data.

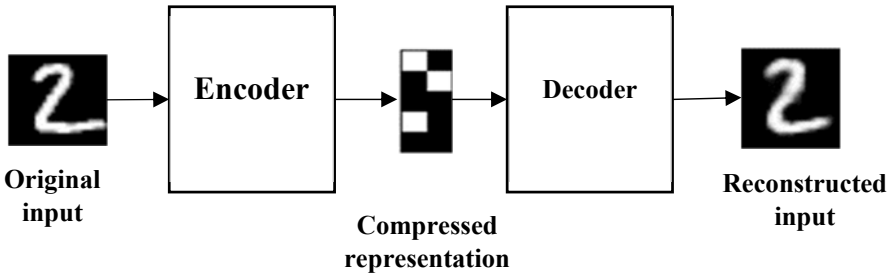

Figure.2. LSTM Compression

3.1 Deep LSTM Neural Network

The LSTM model generally depends on sequence to sequence pattern[31]. The correct output probability is maximised by this trained model resulted in the input sequence. For every training pair (x,y), the parametric model is learned with θ-parameter, which solves the optimisation problem,

$$\theta^* = arg \max_{\theta} \sum_{X,Y} \log p\ (Y\ |X;\ \theta) \qquad (1)$$

From Eq. (1), the overall training is presumed as a sum. A similar architecture is utilised for modelling the probability 'p'. LSTM is designed in such a way as to avoid disappearing gradients, and some long-distance dependencies is remembered from the input sequence. The general LSTM architecture is presented in figure.3. The input is fed in the reversed way, and it has performing empirically well. Through the initial way over input, the LSTM network has expected in learning the input sentences in compact and distributed representation, and the right predictions are start generated. After that, the model proceeds. The following Eq. (2) decomposed by chain rule,

$$p\ (Y|X;\theta) = \prod_{t=1}^{T} p(Y_t\ |Y_1, \ldots\ldots, Y_{t-1}, X; \theta) \qquad (2)$$

No independence assumptions have been made. If the θ^* optimum value is constructed, the DNA sequence compression is estimated through,

$$Y = arg\,\widehat{\max_{Y}}\, p(Y|X;\theta^*) \quad (3)$$

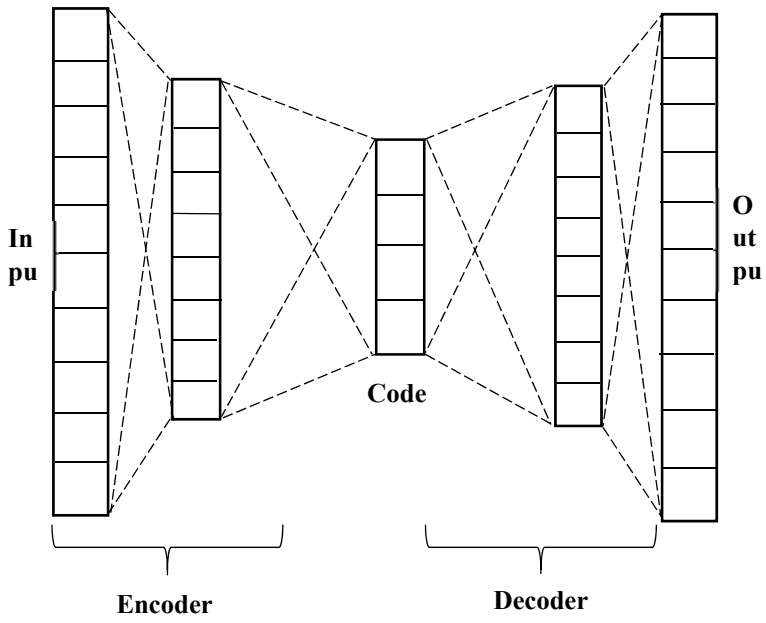

Figure.3. Deep LSTM Architecture

In LSTM model, the sequence to sequence are reviewed. Schmidhuber and Hochreiter [32] modelled the LSTM as described below. Let consider input, control and memory state asa_t, c_t and s_t at a time t. The input sequence is $(a_1, \ldots\ldots, a_t)$, $c$sequence is computed by LSTM as $(c_1, \ldots\ldots, c_t)$ and the s-sequence is $(s_1, \ldots\ldots, s_t)$ are illustrated as,

$$i_t = sigm\,(W_1 a_t + W_2 c_{t-1}) \quad (4)$$

$$i_t{}' = \tan h\ (W_3 a_t + W_4 c_{t-1}) \quad (5)$$

$$f_t = sigm\,(W_5 a_t + W_6 c_{t-1}) \quad (6)$$

$$o_t = sigm\,(W_7 a_t + W_8 c_{t-1}) \quad (7)$$

$$s_t = s_{t-1} \odot f_t + i_t \odot i_t{}' \quad (8)$$

From the above Eq. (8), $\odot$ defines multiplication of element-wise with $(W_1, \ldots\ldots, W_t)$ and c_o are considered as model parameters. Element-wise computation is seen to all non-linearity. The training objective is maximised by the stochastic gradient descent based on all parameters.

This study utilised LSTM architecture. Commonly used stacked LSTM layers allowed the higher layers for higher-order representations learning related to the input. The overfitting is prevented by dense layers. The encoder and decoder process with respect to compression ratio using LSTM is depicted in figure.3., After the GO symbol read, the output layer predicts, three labels are read as 0-if the word is removed, 1-if the word is sustained in compression. The proposed algorithm for the LSTM Neural network with encoder and decoder process is shown below,

Algorithm-LSTM Neural Network with Encoder and Decoder

1. **Function ENCODE, DECODE(X)**
2. Lstm ← CREATE LSTM
3. LayersStates ← INITIALIZELAYERS (Lstm)
4. **for all** $Z_j \in REVERSE(z)$ **do**
5. Layersstate ← ONESTEP(Lstm, LayersState, Z_j)
6. **end for**
7. LayersState ←ONESTEP (Lstm, LayersState, GO)
8. Create the vector. Each item contains the state of the layers, the labels predicted so far, and probability.
9. Encoder←{($Layersstate$, (),1.0)}
10. **for** all $Z_j \in Z$**do**
11. Decode← {}12
12. **for all** ($layersstate, labels, prob$) ∈ $beam$**, do**
    a. ($nextlayersstate, outputs$) ← ONESTEP(Lstm, LayersState, Z_j)
13. **for all** $output \in outputs$**do**
14. $nextbeam \leftarrow nextvector \cup \{(Nextlayerstate, labels + outout\ label, prob * output.prob)\}$
15. **end for**
16. **end for**
17. Decoder← topN (Nextvector)
18. **end for**
19. **return** top (decode)
20. **end function**

In the decoding process, searching through all likely output sequences provided as X by considering Eq. (3). For LSTM, the old history is assessed for every prediction. Eq. (2) cannot be simplified using the Markov supposition. At decoding time, the search space is exponential on input length.

IV. RESULTS AND DISCUSSION

This section discusses the results obtained by implementing the Deep LSTM Neural Network for DNA sequence compression. The proposed system is analysed by comparing it with the existing methods with respect to ratio, time and loss. The analytical results are discussed here.

The below table-1 shows the comparative analysis of the proposed and existing Fast Reference Free method and AGIC in terms of ratio, time and loss.

Table-1. Comparative analysis of the proposed And existing system[33] in terms of various parameters

Algorithm	**Ratio**	**Time**	**Loss**
Fast Reference Free Method	5	42.62 sec	0.2101
AGIC	9	8403.32 sec	0.2107
Proposed	5	2.56	0.156

From the above table-1, it is clear that the existing Fast Reference Free method shows 5 as compression ratio, and the existing AGIC shows 9 as compression ratio. On the contrary, the proposed method exhibits 5 as compression ratio. Here AGIC is found to show a high ratio. But, the optimal compression ratio is 5 as it shows a minimum loss. Moreover, the existing AGIC shows a high loss at a rate of 0.2107. This is followed by the existing Fast Reference Free method, which shows loss at a rate of 0.2101, and the proposed method shows 0.156 as the loss rate. Thus, the proposed method shows minimum loss when compared to the existing methods. Similarly, the existing Fast Reference Free method takes 42.62 sec for compression, the existing AGIC takes 8403.32 sec, and the proposed method takes 2.56 sec to compress the DNA sequence data. Thus the proposed method takes minimum time for compression.

The analysis of the proposed and existing system in terms of ratio is shown in the below figure.4.

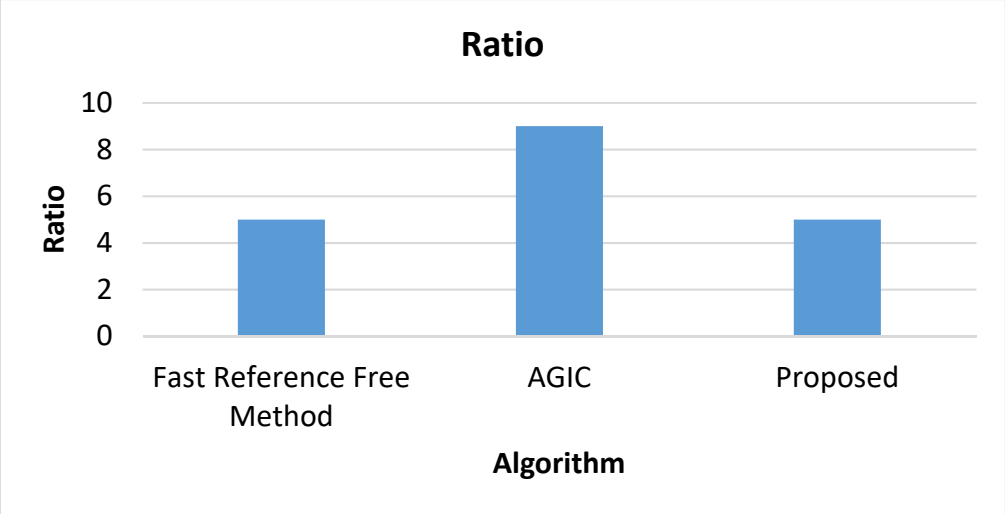

Figure.4. Comparative analysis of the proposed and existing methods [33]in terms of ratio

From the above figure.4, it is clear that the proposed method shows an optimal compression ratio than the existing methods. Here the existing AGIC method is found to show a high compression ratio. But, if the compression ratio exceeds beyond 5 then the loss rate will start increasing. Thus the optimal compression rate is 5, which is attained by the proposed system. Here the existing Fast Reference Free Method also shows 5 as compression rate. But, this method shows high loss when compared to the proposed system. It is graphically shown in the below figure.5.

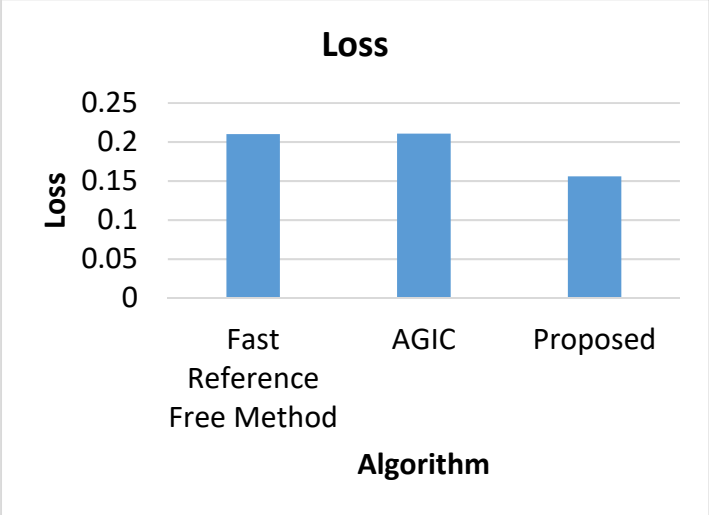

Figure.5. Comparative analysis of the proposedand existing methods [33]in terms of loss

From the above figure.5, it is found that the existing AGIC shows a high loss rate. This is followed by the existing Fast Reference Free Method that exhibits high loss rate than the proposed system. Thus the proposed method shows minimum loss than the other two existing methods.
Hence, the proposed method is found to be effective than the other two existing methods in terms of ratio, time and loss, thereby effectively compressing the DNA sequence data.
Moreover, each DNA sequences (A, G, C and T) are encoded in binary form. It is shown in the below table-2.

Table-2. Binary encoding for DNA sequences

Sequence	Binary Encoded
A	10
G	11
C	101
T	111

The four DNA sequences are encoded in binary. The DNA sequence 'A' is encoded in binary form as 10, the DNA sequence 'G' is encoded as 11, 'C' is encoded as 101, and 'T' is encoded as 111. In this way, the four DNA sequences are encoded and later processed. In addition, the original size of the sequence before compression and after compression is computed, and the results of the proposed and existing methods are comparatively analysed, as shown in the below table-3.

Table-3. Comparative analysis[11] in terms of compression ratio

		Compression ratio percentage (%) after compression			
Dataset	**The original size of sequence before compression (Bytes)**	**BDNAS algorithm**	**Lossless compression**	**CompressBest algorithm**	**Proposed**
Human Gene	58,864	33.33	32.61	32.53	24.56

From the above table-3, it is found that the original size of the sequence before compression is 58,864 bytes. Thus, after compression, the existing algorithms shows maximum compression rate when compared to the proposed system. It is graphically shown in the below figure.6.

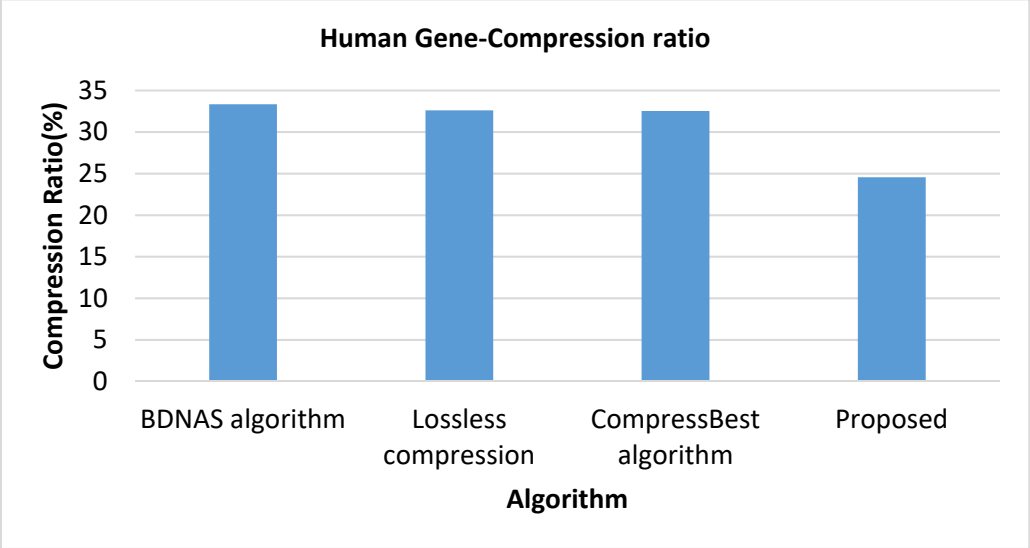

Figure.6. The compression ratio of DNA sequences[11]

From the above figure.6, it is found that the proposed system is effective than the existing system in terms of compression rate as it exhibits minimum compression rate when compared to the existing method.

Subsequently, the count of each DNA sequences are determined, and it is graphically shown in the below figure.7.

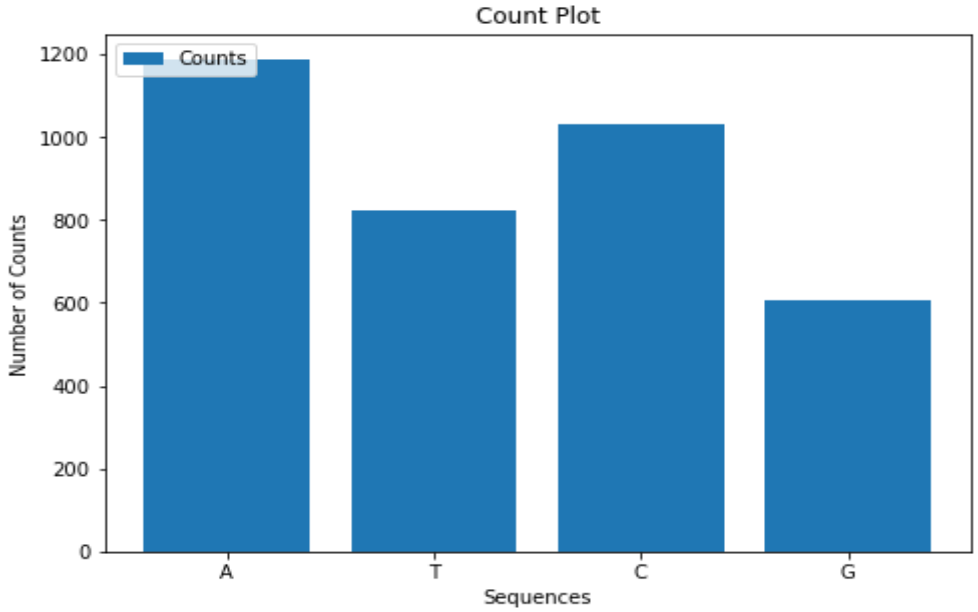

Figure.7. Count of DNA sequences

From the above figure.7, it is clear that the count of DNA sequence 'A' is 1200. The count of DNA sequence 'T' is 800, the count of DNA sequence 'C' is 1000, and the count of DNA sequence 'G' is 600. Hence, the count of DNA sequence 'A' is higher, followed by 'C', 'T' and 'G'.

V. CONCLUSION

DNA sequence compression is vital in the research area as it has various merits. Thus, the vectorizer technique is performed using the encoder, channel and decoder. Then, the encoder and decoder are set for compression. This study proposed a Deep LSTM Neural Network to compress the DNA sequence data. The compression results are analysed in terms of various metrics to evaluate the efficiency of the proposed system. The metrics include ratio, time and loss. The analytical results revealed that the proposed system exhibited minimum compression loss, optimal compression ratio and utilised the minimum time to compress the DNA sequence data. These results revealed the efficiency of the proposed system in compressing the DNA sequence. Additionally, the performance of the proposed method is analysed by comparing it with the traditional methodologies. The results from comparative analysis also revealed the efficacy of the proposed methodology.

Conflicts of interest

"The authors have no conflicts of interest to declare"

REFERENCES

[1] P. Mishra, C. Bhaya, A. K. Pal, and A. K. Singh, "Compressed DNA Coding Using Minimum Variance Huffman Tree," *IEEE Communications Letters,* vol. 24, pp. 1602-1606, 2020.
[2] A. Keerthy and S. M. Priya, "Lempel-ziv-welch compression of dna sequence data with indexed multiple dictionaries," *Int. J. Appl. Eng. Res,* vol. 12, pp. 5610-5615, 2017.
[3] B. Carpentieri, "Compression of Next-Generation Sequencing Data and of DNA Digital Files," *Algorithms,* vol. 13, p. 151, 2020.
[4] S. Ranjitha and L. Robert, "Enhanced Tunneled BWT with Longest Common Prefix Array for largest DNA Sequence Compression."
[5] D. Mansouri, X. Yuan, and A. Saidani, "A new lossless dna compression algorithm based on a single-block encoding scheme," *Algorithms,* vol. 13, p. 99, 2020.
[6] B. Saada and J. Zhang, "DNA sequence compression technique based on nucleotides occurrence," in *Int. Multi-Conf. of Engineers and Computer Scientists, Hong Kong*, 2018.
[7] A. Neha and A. Salim, "Towards Context-Aware DNA Sequence Compression Algorithms," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019, pp. 1-4.
[8] D. Pratas, M. Hosseini, J. M. Silva, and A. J. Pinho, "A reference-free lossless compression algorithm for DNA sequences using a competitive prediction of two classes of weighted models," *Entropy,* vol. 21, p. 1074, 2019.
[9] A. Al-Okaily, B. Almarri, S. Al Yami, and C.-H. Huang, "Toward a better compression for DNA sequences using Huffman encoding," *Journal of Computational Biology,* vol. 24, pp. 280-288, 2017.
[10] D. Pratas and A. J. Pinho, "A DNA sequence corpus for compression benchmark," in *International Conference on Practical Applications of Computational Biology & Bioinformatics*, 2018, pp. 208-215.
[11] K. Punitha and A. Murugan, "A novel algorithm for DNA sequence compression," in *Emerging Research in Computing, Information, Communication and Applications*, ed: Springer, 2019, pp. 151-159.

[12] J. Lázaro-Guevara and K. Garrido, "A New approximate matching compression algorithm for DNA sequences," *bioRxiv,* p. 853358, 2019.
[13] M. Silva, D. Pratas, and A. J. Pinho, "Efficient DNA sequence compression with neural networks," *GigaScience,* vol. 9, p. giaa119, 2020.
[14] S. Chowdary, S. V. Kumar, D. Nedunuri, and V. Gupta, "A novel approach to compress dna repetative sequences in bio-informatics," in *Journal of Physics: Conference Series*, 2019, p. 012026.
[15] K. Banerjee and V. Bali, "Design and Development of Bioinformatics Feature Based DNA Sequence Data Compression Algorithm," *EAI Endorsed Trans. Pervasive Health Technol.,* vol. 5, p. e5, 2020.
[16] J. Partee, R. Hazell, A. Solsi, and J. Santerre, "Compressed DNA Representation for Efficient AMR Classification," *SMU Data Science Review,* vol. 3, p. 5, 2020.
[17] R. Wang, Y. Bai, Y.-S. Chu, Z. Wang, Y. Wang, M. Sun, *et al.*, "DeepDNA: A hybrid convolutional and recurrent neural network for compressing human mitochondrial genomes," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 270-274.
[18] S. Bibi, J. Iqbal, A. Iftekhar, and M. Hassan, "Analysis of Compression Techniques for DNA Sequence Data," *arXiv preprint arXiv:2006.02232,* 2020.
[19] A. Womack, "Cigarcoil: A new algorithm for the compression of dna sequencing data," 2019.
[20] S. E. Zahra, K. Masood, and M. Asif, "DNA Compression using an innovative Index based Coding Algorithm," in *2019 22nd International Multitopic Conference (INMIC)*, 2019, pp. 1-6.
[21] M. Goyal, K. Tatwawadi, S. Chandak, and I. Ochoa, "Deepzip: Lossless data compression using recurrent neural networks," *arXiv preprint arXiv:1811.08162,* 2018.
[22] S. M. Hossein, D. De, P. K. D. Mohapatra, S. P. Mondal, A. Ahmadian, F. Ghaemi, *et al.*, "DNA Sequences Compression by GP² R and Selective Encryption Using Modified RSA Technique," *IEEE Access,* vol. 8, pp. 76880-76895, 2020.
[23] S. M. Hossein, A. Mitra, P. K. D. Mohapatra, and D. De, "A Compression & Encryption Algorithms on DNA Sequences Using R 2 P & Selective Technique."
[24] G. P. Arya, R. Bharti, D. Prasad, and V. Garg, "An improved method for DNA sequence compression," in *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, 2017, pp. 1-4.
[25] A. B. Majumder and S. Gupta, "Dynamic Programming Based DNA Compression Algorithm through Substitution Method," *International Journal of Engineering and Applied Sciences,* vol. 4, p. 257356.
[26] K. Punitha and A. Murugan, "SegmentationBased DNA Sequence Compression."
[27] F. Kounelis and C. Makris, "Comparison between text compression algorithms in biological sequences," *Information and Computation,* vol. 270, p. 104466, 2020.
[28] T. Paridaens, "Compression and interoperable representation of genomic information," Ghent University, 2018.
[29] C. Yin, "Encoding and decoding DNA sequences by integer chaos game representation," *Journal of Computational Biology,* vol. 26, pp. 143-151, 2019.
[30] K. Kryukov, M. T. Ueda, S. Nakagawa, and T. Imanishi, "Nucleotide Archival Format (NAF) enables efficient lossless reference-free compression of DNA sequences," *Bioinformatics,* vol. 35, pp. 3826-3828, 2019.
[31] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *arXiv preprint arXiv:1409.3215,* 2014.
[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, pp. 1735-1780, 1997.
[33] T. Islam, C. H. Kim, H. Iwata, H. Shimono, A. Kimura, H. Zaw, *et al.*, "A Deep Learning Method to Impute Missing Values and Compress Genome-wide Polymorphism Data in Rice," 2021.

Authors Profile

Prashanth.M.C author is currently pursuing Ph.D. in the department of computer science, Kuvempu University, Karnataka, INDIA. He is working in the domain of DNA sequences pattern matching. This author has completed his Master degree in computer science in 2013.

M.Ravikumar the author has completed his B. E, M. Tech and Ph.D. degrees in the year 1996,2001 and 2016 respectively. His Ph.D. topic is Estimationof Multiple Skews in Trilingual Handwritten Document Images. His research area includes Document images analysis, pattern matching, medical image processing. He has published more than fifteen papers and journals.