International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 1, Jan-Feb-2021, ISSN (Online): 2395-566X

A Review on Lossless Data Compression Techniques

Rahul Barman, Sharvari Deshpande, Prof. Dr. Nilima Kulkarni, Shruti Agarwal, Sayali Badade
Department of Computer Science and Engineering
MIT ADT University
Pune, India
rahulbarman1215@gmail.com, sharvarideshpande24@gmail.com, nilima.amrita@gmail.com,
shrutileena@gmail.com, sayalib44@gmail.com

***Abstract*- Data is being generated at an exponential rate in every sector of the world. This much amount of data can prove to be very costly in terms of data storage, transfer speed and infrastructure. Due to this many industries have shifted their focus on cloud-based storage technology. Lossless data compression is one of the important approaches to solve the problem of excessive storage consumption, maintaining data integrity, transfer efficiency and achieving higher streaming speed. Lossless data compression refers to the process of modifying and converting the bit structure of data such that it consumes less storage space and provides near loss of original data. There exists different algorithms and techniques for performing the compression on different types of data formats. This paper identifies different lossless data compression techniques in existence and a conclusion is drawn based on these identified methods. Comparative analysis has been made of different algorithms used in various referenced papers. It describes the future scope and application of compression algorithms in several fields.**

***Keywords*- Data Compression, Lossless, Adaptive, Machine Learning, Huffman Encoding, Delta Encoding.**

I. INTRODUCTION

In today's world, digitalization is at its peak. According to the survey, over 2.5 quintillion bytes of data are generated every day. The amount of data created every year is growing faster than ever before. By the end of this year, every human will be creating 17 megabytes of information each second. Due to such rapid development in digitalization which contributes in the rapid growth of data, there will be more storage required for it which implies an increase in the cost for storage. To solve this problem, Data Compression is one of the most important approaches. It is a process of reducing the number of bits required to express data in a compact form.

Data Compression consists of two main classes, one is Lossless Compression and other one is Lossy Compression. Lossless data compression reconstructs the original data from the compressed data. In lossless compression, there is no loss in the quality of the data and every bit regains its original value after the decompression that maintains data integrity. Lossy compression as the name suggests, results in the loss of some amount of data. In lossy compression, the data eliminated is not noticeable and data quality is compromised.

The data compression can be done on various data formats namely text, image, video, audio etc. There are different compression techniques available used for different applications. For the text compression, Dictionary Based Coding, Run Length Coding, Statistical Based techniques and Entropy Coders are used. Huffman Encoding and Lempel-Ziv-Welch algorithms are found to be more efficient in the text format for the data compression. DCT

(Discrete Cosine Transformation),DWT (Discrete Wavelet Transform), JPEG (Joint Photographic Expert Group), Delta Encoding, LZW are some of the techniques used for image/video compression. As the compressed file requires significantly less storage capacity than the uncompressed file so the compression of data results in saving the storage capacity which implies a decrease in the cost required for that storage which is the ultimate goal. Also, Data Compression improves the transfer efficiency and helps to achieve higher streaming speed even with lower bandwidth.

II. LITERATURE SURVEY

As described by **Komal Sharma & Kunal Gupta in the paper [1] "Lossless Data Compression Techniques and their performance", 2017**, there are various lossless data compression techniques. The paper describes these techniques and uses time complexity and space complexity to compare their performance and efficiency. The use of lossless data compression techniques is done on several media file formats (txt, pdf, jpg, etc). Huffman Coding lowers the number of bits for binary codes which occurs frequently whereas LZW works by reading the sequence of symbols. To generate binary code based on symbol frequency Shannon-Fano coding is used.

Thus, they calculated the mean compression ratio (CR), compression factor (CF) and compression time (CT) for various media files and analysed the result to select the

most suited method of compressing any file format. From the results and analysis, it was concluded that for text files, Huffman algorithm is found to be the best for compression as it shows the minimum CR value and takes optimal compression time. LZW compression algorithm works well for image file formats while Shannon-Fano algorithm works well for hybrid text file and audio file formats. Though this paper describes various classical methods used for compression of several file formats, ML model has not been used for implementation which makes the model non-adaptive.

In the paper [2] "A fast and efficient lossless Data Compression method", 1999, authors proposed a lossless data compression algorithm using arithmetic coding method. The paper employs a fuzzy-tuning modeler which is adaptive and that applies fuzzy inference concepts to efficiently deal with the problem of conditional probability estimation. When compared to other lossless coding schemes, the compression results of the above method are good and satisfactory for various types of data. A table-lookup approach has been adopted for the fuzzy-tuning modeler which makes the system fast, simple and suitable for VLSI implementation. The main drawback of this method is that it uses single bit multiplication of each input value to get a more accurate and precise result. However, it can achieve high encoding speed by using a simple parallel multiplier.

Another paper [3] entitled "Comparison of Lossless Data Compression techniques", 2020, shows a complete study of various lossless compression techniques. The paper aims to evaluate the best compression algorithm based on the space-saving, execution time needed, compression ratio and rate of already existing lossless compression techniques for the storage and communication. The analysis shows that the LZW compression technique outperforms the other lossless compression techniques discussed. It results in better performance and gives a good compression ratio of 4:1 with 76.9% space saving. Also, it will not take large bandwidth to send the massive amount of data from the transmitter to the receiver. As compared to other classical compression algorithms Delta Encoding generally takes less compression and shows better performance.

The paper [4] "Lossless Compression of Colour Map Images by Context Tree Modelling", 2007, have proposed a method for lossless compression of colour map images by splitting the colour maps into different layers and compressing the binary layers individually using an optimized context tree model which makes use of interlayer dependencies. It aims to obtain a better result by working directly on the colour image without layer separation. The method extends the previous context tree based method to work on colour values instead of binary layers. First, an n-ary context tree is generated by constructing a complete tree upto a predefined depth and then the nodes that do not provide compression improvements are pruned out. The pruning algorithm decreases the amount of time needed in the optimization of the tree structure. The experiments shows that the stated method gives better performance than the existing methods by 20% for a large set of different colour map images. The compression method was successfully applied to raster map images upto 67 colours. The further scope of study is to apply this algorithm to photographic images. Though this algorithm works well, it can be used only for images and not for other types of file formats which can be a limitation. Also, it does not make use of the ML model.

Komal Sharma and Kunal Gupta [5], 2017, have proposed their own compression technique and compared its performance with Huffman Encoding Algorithm. The main approach is to compress the original data in a lossless manner. The technique employs non-prefix codes of static variable length for ASCII characters assigned by analysing the probability of character occurrences in lossless data files.The input filename, encryption key and the assigned codes are used for parsing and encoding the input file. For comparing and analyzing the algorithm four different methods are used which are, CR (compression ratio), CT (Compression Time), CF (Compression Factor) and SP (saving percentage). They have worked on various data formats.

For decompression, characters in compressed files are deciphered using encryption keys and unwrapped as their respective bitcodes. Results presented in the paper show equivalent statistics for the proposed algorithm and Huffman Encoding.It can be analysed that for compressing text files, image files and video files the algorithm described in the paper is suited the best. The stated technique makes use of static bitcodes making it non adaptable in nature, which is the limitation of this technique. As stated in the paper, future work can be on a similar technique which takes less compression time and has assigned codes modifiable.

In the paper [6] by Raghuvendra Pratap Tripathi and G.R.Mishra(2017), authors have researched and analysed different lossless compression techniques. They have presented the analysis of time domain techniques like Amplitude zone time epoch coding (AZTEC), Turning Point (TP) technique and techniques based on transformation like Discrete Cosine Transform (DCT) technique and Empirical Mode Decomposition (EMD) technique. The analysis of performance is judged on the basis of Compression Ratio (CR) and Percent Root Mean Squared Difference (PRD).

The paper focuses on an efficient technique for performing compression of recorded ECG signals. ECG signals are the real time data recorded in the Tele-cardiology procedures. The results produced in the paper indicated CR and PRD values for all the four lossless compression techniques and

compared them with each other. It has been inferred that the AZTEC technique provides a smaller value of CR with maximum PRD value, however the opposite result is found in case of EMD technique.

Karaca, A.C. and Gullu, M. K. [7] ,2017, have described a method called recursive least square (RLS) method for lossless compression and decompression of ultraspectral sounder dataset in their paper. The ultraspectral dataset consists of 3D data related to atmospheric changes, wind profiles, and moisture flux. It is generated using optical instruments like a grating spectrometer or an interferometer. The RLS algorithm has been elaborated which consists of seven steps. At the initial step, it predicts the pixel values using weight vectors for each spectral band. The prediction error weight calculations are being updated each time and the iteration will be performed for each and every pixel. The paper highlights the importance of selecting vector length used in the method. The vector length in band prediction directly affects the compression ratio. The results of the experiments done by the authors provide comparable compression ratios of the RLS method to the two methods ADQPCA and FSQPCA, which has outperformed the other two methods. It can be concluded that the given method can be implemented in embedded systems.

As explained by Xiwen Zhao and Zhihai He in the paper [8] entitled "Local Structure Learning and Prediction for Efficient Lossless Image Compression", encoding high-frequency structures like edges, contours, and texture regions in an image is one of the big challenges. To overcome this challenge, the authors have developed an encoder for images based on structure prediction. The image is first classified into structured and unstructured regions. The unstructured regions are encoded with existing image compression techniques like CALIC and structure regions with structure prediction. The experimental results in the paper shows that the scheme proposed is efficient in performing lossless image compression on images with significant structured components. In future, an algorithm can be introduced to determine the optimal classification threshold in terms of efficiency in coding.

As described by N.Sriraam, R.Kannan and C.Eswaran in the paper [9] "Lossless Compression of EEG data using Neural Network Predictors", 2002, the two-stage compression method has been stated. The two-stage lossless compression technique involves Neural Network predictor in the first stage and the traditional compression algorithms in the second stage. The method uses the EEG (Electroencephalography) data in the form of signals for the compression in a lossless manner. In the first stage of Neural Network predictor, MLP (Multi-Layer perceptron) and EN (Elman Network) are used.

The two models namely Feedforward and Feedback one with activation functions, linear and sigmoid are used in neural networks. The EEG data signals were initially converted into ASCII code. Fixed training and Block Adaptive training are the two training processes used while in fixed training, the entire input represented as a single block and used to train the neural network. In the block adaptive training, the input is divided into various blocks and then neural networks trained for each block. The compression is achieved as the residue data size is much smaller than the data size of the input. Compression ratio being an evaluation criterion shows that the MLP model works better than the EN model and also the result shows that the block adaptive training gives better performance than the fixed training. In the second stage of the method, the traditional entropy coding methods namely Huffman Encoding and Arithmetic Encoding have been used for achieving a better performance of lossless compression. The method stated above only used for the EEG data format and the overall computational complexity is much higher than the traditional techniques.

In the paper [10], "Evaluation of Lossless Compression Techniques", 2015, Smitha Rao and Pratima Bhat discussed multiple challenges in compression of different data formats, different compression techniques used for the various data formats and about the evaluation criteria for the different compression techniques. The main focus for this study is to evaluate different compression techniques available in various applications where the data integrity is maintained with high quality. Compression Efficiency, Resolution, Error Resiliency, Power Consumption,

Data Rates are the various challenges in the compression discussed in detail. To measure the performance of the data compression, evaluation criteria like Compression Ratio, Image Quality which includes PSNR and MSE, the Compression Rate, Distortion, Coding complexity and Compression Speed, Latency were mentioned in the paper at a good knowledge. Different compression techniques for various data formats like text, video and images, audio etc were discussed where in the text compression, two methods namely Dictionary based and Statistical based were mentioned. After the comparison, In Dictionary based LZW performs better than the LZ77 and LZ78 algorithms. In statistical based technique, Arithmetic Coding gives the best lossless compression. For image compression, In DWT based compression technique JPEG-LS gives better performance than lossless JPEG method as per image quality and compression ratio.

The paper named [11] "Lossless Compression for Volumetric Medical Images Using Deep Neural Network with Local Sampling" by O. H. Nagoor, J. Whittle, J. Deng, B. Mora, M. W. Jones explained a new lossless compression algorithm for 16-bit image volumes. The method comprises training Neural Network as 3D data predictor which reduces the loss with the original data and compresses the output obtained using arithmetic

coding. The importance of lossless compression in Medical Imaging which is used in the section of clinical diagnosis as any loss in the image data is not expected and also the importance of high quality and high resolution in the medical imaging are mentioned. The evaluation based on performance of compression in bbp (bits per pixel) of the stated neural network model in comparison with the best latest lossless compression techniques are done which shows the proposed system gives the higher compression ratio than the JPEG-LS, JP3D,JPEG2000 and HEVC. The dataset used are the CT scans of patients which are stored in 16-bit grayscale images. The method only uses the volumetric image data.

In the paper [12] entitled "Data Compression through Adaptive Huffman Coding Scheme", 2000, describes the system for data compression which uses adaptive Huffman Coding approach for generating variable-length-code. The system consists of encoder and decoder processes. Encoder process used to compress the input text character by character, characters being encoded by variable-length-code. Frequently used characters are assigned by shorter bit codes and the infrequently occurred characters are represented by longer bit codes.

The adaptive approach automatically gets updates according to the change in the text data. The decoder process outputs the encoded text back to original text. Re-scaling, sibling, escape error, swapping and the important one adaptive are the approaches in the system used to improve the performance of the compression. The proposed system works only on the text data formats.

As discussed in the paper [13] "A lossless compression algorithm for video frames", 2014, The current method of compression for video frames occurs in blocks. The paper suggests using line by line pixel compression using dictionary and non-dictionary pixels along with Huffman coding. The code words are converted into bit streams which are then compressed using a predefined set of rules. The average data reduction ratio obtained by the system was from 19.77% to 54.47%. The reason behind the idea of using line by line compression of video files was to reduce computational power and delay. The downside of using the technique is that line by line compression makes the system very slow to compress the whole video file. Also, the system can only compress video files and the compression ratio varies to a large range.

The paper [14] entitled, "DeepZiv: Lossless Data Compression using Recurrent Neural Networks", 2019, describes that the use of Machine Learning greatly increases the performance barrier for compression algorithms. The paper emphasised the importance of sequences in text and geospatial data and how linking these sequences increases the compression ratio. The paper discussed using RNN (Recurrent Neural Networks) for detecting the sequences and using the sequence to predict a compressed sequence. The decompressor takes this compressed sequence and predicts the original sequence. This system achieves 20% better compression ratio then Gzip compression algorithm. However, sequences are mainly found on text data, therefore cannot be used as an efficient technique for compression of other formats like images, audio, video etc.

Data compression in time series databases has risen in importance a considerable amount as depicted in the paper [15] "Two-Level Data Compression using Machine Learning in Time Series Database", 2020. This paper shows the importance of considering the local context around individual points. The paper depicts a system which uses a two-level compression model that selects a proper compressions scheme for each individual point, so that diverse patterns can be captured at a fine granularity. The proposed system improves compression ratio upto 120% (with an average of 50%) compared to other time series compression models. But the main downside of this system is that it can only be used for compressing time series data. Data formats like text, images, videos may not necessarily contain time series data and hence this method fails for other data formats.

Ren´e J.v an der Vleuten, and Philips in their paper [16] "Low-Complexity Lossless and Fine-Granularity Scalable Near-Lossless Compression of Color Images", 2015, explained about using the LSB (Least Significant Bit) and MSB (Most Significant Bit) for compression of pixel values. The MSB is compressed in lossless manner by a dedicated compression method and LSB is compressed by a scalable encoder. The system also suggested applying color transform in a lossless manner to recolor the image before compressing. This improves the compression ratio by a significant amount. The main disadvantage of using the algorithm described is that it may sometimes fail to split the point values which may result in lower compression ratio. Hence for complex data points this system fails to be efficient.

Merits and demerits with results of different algorithms from studied paper are discussed in Table 2 as follows:

Algorithms	Merits	Demerits

In the paper by Komal Sharma & Kunal Gupta, various classical techniques such as Huffman coding, Lempel Ziv and Shannon-Fano Coding have been used.	Performance and efficiency have been compared by using both space and time complexity. Also, it makes use of most of the file formats. Huffman coding for txt, LZW for bmp, png, gif, Shannon-Fano Coding for doc, pdf, docx, wav.	Machine Learning model has not been used in the implementation part thus making it non-adaptive in nature. The data can be compressed only upto a specific threshold value and thus it can't be compressed further.
As described in the paper "A fast and efficient lossless data compression method" arithmetic coding algorithm has been used for online lossless compression of data.	The algorithm gives good efficiency in terms of compression. It employs an adaptive fuzzy-tuning modeler which applies the fuzzy inference concept to deal with the problem of conditional probability estimation efficiently.	It uses single bit multiplication of each input to get a more accurate result. The system requires hardware implementation too.
In the paper by Raghuvendra Pratap Tripathi and G.R.Mishra, four different lossless data compression techniques such as Amplitude zone time epoch coding (AZTEC), Turning Point (TP), Discrete Cosine Transform (DCT) and Empirical Mode Decomposition (EMD).	AZTEC techniques provided better results as compared to EMD techniques, in terms of CR and PRD values.	Specifically, for Tele-cardiology procedures to perform compression on ECG signal real time data.
Komal Sharma and Kunal Gupta in their paper have employed a new compression technique called H_C algorithm. It uses encryption along with it.	For compressing text files, image files, and video files, H_C algorithm is appropriate. It shows the least compression time for above file formats. Encryption ensures data security.	Assigned codes are non-modifiable.

III. CONCLUSION

This paper presents a comparative study on various existing techniques and algorithms for data compression. It has been observed that most of the methods provide efficient compression for only specific data formats. Thus, a new method for adaptive lossless data compression and decompression on various data formats like text (.txt, .doc, .docx) and images (.png, .jpg, .jpeg) using predefined and machine learning algorithms can be proposed. It can be made adaptive such that the compression ratio keeps getting better over time and gives good results as compared to the classical lossless data compression techniques. Also, use of machine learning technique will ensure that different file formats can be compressed within the same system. The future scope is to ensure that all types of file formats including the hybrid files, can be compressed with high compression ratio and lower compression time.

IV. ACKNOWLEDGMENT

I express my profound thanks to my Guide Dr. Nilima Kulkarni for her expert guidance, encouragement and inspiration during this project work.

I would like to thank Dr.Reena Pagare, Project Coordinator,Department Computer Science & Engineering for extending all support during the execution of the project work.

I sincerely thank Prof. Dr. Rajneesh kaur Sachdeo, Head, Department of Computer Science & Engineering, MIT School of Engineering, MIT-ADT University, Pune, for providing necessary facilities in completing the project.

REFERENCES

[1] K. Sharma and K. Gupta, "Lossless data compression techniques and their performance," International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 2017, pp. 256-261, doi: 10.1109/CCAA.2017.8229810. (2017)

[2] Jer Min Jou and Pei-Yin Chen, "A fast and efficient lossless data-compression method," in IEEE Transactions on Communications, vol. 47, no. 9, pp. 1278-1283, Sept. 1999, doi: 10.1109/26.789659.

[3] A. Gopinath and M. Ravisankar, "Comparison of Lossless Data Compression Techniques," International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2020, pp. 628-633, doi: 10.1109/ICICT48043.2020.9112516. (2020)

[4] Akimov, A. Kolesnikov and P. Franti, "Lossless Compression of Color Map Images by Context Tree Modeling," in IEEE Transactions on Image Processing, vol. 16, no. 1, pp. 114-120, Jan. 2007, doi: 10.1109/TIP.2006.887721.

[5] K. Sharma and K. Gupta, "Lossless data compression technique with encryption based approach," 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-5, doi: 10.1109/ICCCNT.2017.8204117. (2017)

[6] R. P. Tripathi and G. R. Mishra, "Study of various data compression techniques used in lossless compression of ECG signals," International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 2017, pp. 1093-1097, doi: 10.1109/CCAA.2017.8229958. (2017)

[7] A. C. Karaca and M. K. Güllü, "Lossless compression of ultraspectral sounder data using recursive least squares," 8th International Conference on Recent Advances in Space Technologies (RAST), Istanbul, 2017, pp. 109-112, doi: 10.1109/ RAST .2017.80 02 966. (2017)

[8] X. Zhao and Z. He, "Local structure learning and prediction for efficient lossless image compression," IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, 2010, pp. 1286-1289, doi: 10.1109/ICASSP.2010.5495420. (2010)

[9] N. Sriraam, R. Kannan and C. Eswaran, "Lossless compression of EEG data using neural network predictors," Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02., Singapore, 2002, pp. 2046-2048 vol.4, doi: 10.1109/ICONIP.2002.1199034.

[10] S. Rao and P. Bhat, "Evaluation of lossless compression techniques," International Conference on Communications and Signal Processing (ICCSP), Melmaruvathur, 2015, pp.1655-1659, doi: 10.1109/ICCSP.2015.7322799. (2015)

[11] O. H. Nagoor, J. Whittle, J. Deng, B. Mora and M. W. Jones, "Lossless Compression for Volumetric Medical Images Using Deep Neural Network with Local Sampling," IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2020, pp. 2815-2819, doi:10.1109/ICIP4 0778.2 020.9191031. (2020)

[12] M. Y. Javed and A. Nadeem, "Data compression through adaptive Huffman coding schemes," TENCON Proceedings. Intelligent Systems and Technologies for the New Millennium (Cat. No. 00CH37119), Kuala Lumpur, Malaysia, 2000, pp. 187-190 vol.2, doi: 10.1109/TENCON.2000.888730. (2000)

[13] K. V. Arya and N. Tato, "A lossless compression algorithm for video frames," 9th International Conference on Industrial and Information Systems (ICIIS), Gwalior, 2014, pp. 1-5, doi: 10.1109/ICIINFS.2014.7036621. (2014)

[14] M. Goyal, K. Tatwawadi, S. Chandak and I. Ochoa, "DeepZip: Lossless Data Compression Using Recurrent Neural Networks," Data Compression Conference (DCC), Snowbird, UT, USA, 2019, pp. 575-575, doi: 10.1109/DCC.2019.00087. (2019)

[15] X. Yu et al., "Two-Level Data Compression using Machine Learning in Time Series Database," IEEE 36th International Conference on Data Engineering (ICDE), Dallas, TX, USA, 2020, pp. 1333-1344, doi: 10.1109/ICDE48307.2020.00119. (2020)

[16] R. J. van der Vleuten, "Low-complexity lossless and fine-granularity scalable near-lossless compression of color images," Proceedings DCC 2002. Data Compression Conference, Snowbird, UT, USA, 2002, pp. 477-, doi: 10.1109/DCC.2002.1000020. (2002)