International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Higher Payload Capacity in DNA Steganography using Balanced Tree Data Structure

Partha Saha, Lubna Yasmin Pinky, Mohammad Ashraful Islam, Papia Akter

Abstract: ***In recent years, with the huge expansion of internet and communication technology, the data transmission rate has increased exponentially. The threat of unauthorized penetration to the secret messages during transmission has become a major concern for data integrity. Cryptography and Steganography are two well-known techniques used to secure and hide confidential data from the intruders. Cryptography is used to obscure the secret message whereas Steganography embeds the messages into a cover media and conceals the presence of secret information. In DNA steganography, DNA molecular sequence is used as a cover medium. In the field of steganography, Payload capacity is the measurement of hiding intended messages in cover media. The Capacity of hiding messages in cover media is one of the prime challenges in the field of steganography. The principal study of this research is to provide a new framework using DNA steganography that provides a higher payload capacity. We have used balanced tree data structures for message encoding where the leaf node contains the intended message. This unique process of message encoding and decoding guarantees a payload capacity of ≥ 0.50.***

Keywords: Balanced Tree, DNA Steganography, Payload Capacity.

I. INTRODUCTION

In this era of science and technology, information sharing and data transfer through the internet has increased abruptly. In recent years, with the huge development of the internet and communication technology, the data transmission rate has increased exponentially. The threat of cybercriminals accessing confidential information and the transfer capacity has been a major concern to the data communication experts. So secure and efficient data transmission has been gained more and more attention, and become a hot topic in the data transmission research field [1]. Cryptography and steganography are the two most popular techniques to overcome this problem.

Cryptography refers to secure and protect confidential information by scrambling its content and converts it into an unreadable format. In cryptography, the secret message of the sender is encrypted using an encryption algorithm with a secret key and then transmit the secret key and encrypted message to the recipients. The recipients decrypt the encrypted message with the help of secret key and appropriate decryption algorithm. An unauthorized user will not be able to extract the secret message without the secret key [2].

Steganography is a technique of hiding the existence of the confidential message into the cover medium [10]. Various steganography techniques have been introduced based on the cover medium. Among these techniques, images steganography, audio steganography, video steganography, text steganography, etc. are the most popular. A lot of methods and algorithms have been introduced on these above steganography techniques. But the hiding capacities of these techniques are very low [3]. To increase the hiding capacity into a cover medium, the DNA steganography is introduced.

The DNA is the Deoxyribonucleic acid and made up of nucleotide which contains a phosphate group, a sugar group, and a nitrogen base. The four different types of nitrogen bases are adenine (A), thymine (T), guanine (G) and cytosine (C). The DNA bases pair up with each other, adenine (A) pair up with thymine (T) and cytosine (C) pair up with guanine (G). In DNA steganography, a DNA sequence is used as a cover medium to hide confidential information [11].

The general idea behind DNA Steganography is to choose a random DNA sequence as a cover medium and then concealing the secret message into it using an encryption algorithm and transfer the modified DNA sequence to the receiver. The receiver extracts the secret message from the modified DNA sequence using the appropriate decryption algorithm, [4, 10].

Most of the existing methods in steganography need a huge size of reference DNA sequence with respect to the secret message, more precisely three or four times larger than the secret message that to need to hide.

In this paper, we propose a new framework of DNA steganography, which increases the payload capacity. We have used a random DNA sequence as cover media. Another DNA sequence is generated by encrypting intended message. A balanced tree data structure is generated with that random DNA sequence. Later to get the reference DNA sequence, all the leaf nodes are replaced by the encrypted DNA (intended message).

Manuscript published on November 30, 2019.
* Correspondence Author
Partha Saha, Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh.
Lubna Yasmin Pinky, Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh.
***Mohammad Ashraful Islam**, Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh.
Papia Akter, Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh.

In our designed balanced tree data structure, the number of the internal node is always ≤ the number of the leaf nodes. So, the reference DNA sequence is not more than two times in size than encrypted.

II. LITERATURE REVIEW

Wang Z, Zhao X, Wang H, Cui G [5] proposed a method on their "Information hiding using DNA steganography" paper where the information is encrypted using vigenere cryptography and decomposed the ciphertext again and again until the ciphertext is not assumed or contaminated and then using the encoding algorithm the resulting vigenere ciphertext is concealed into DNA sequence[10].

S. Jiao, and R. Goutte [6] proposed a technique on paper "Hiding data inside the DNA of a living organisms" where common bacteria are used for encryption. Bacillus subtilis gene (tatAD) has been selected for the reference DNA sequence. The message is translated into binary and converted to DNA bases of codons. Then some codons of reference DNA sequence are replaced by the codons of the secret message using the properties of silent mutation. In the encryption process, any encryption algorithm such as DES, AES, RSA, Blowfish, etc. can be used for additional security.

Khalifa A., [7] proposed an algorithm called LSBase (Least Significant Base Substitution) on paper "LSBase: A key encapsulation scheme to improve hybrid crypto-systems using DNA steganography" to improve the security of the secret message. It uses the key encapsulation mechanism of a hybrid cryptosystem for better key management and hides the confidential message into DNA sequence using the codon degeneracy without modifying the functionality of the DNA sequence. To improve security, it converts the DNA sequence into an RNA sequence. The decryption process is done blindly without the help of the reference DNA sequence.

Malathi P, Manoaj M, Manoj R, VaikunthRaghavan, Vinodhini R. E, [8] proposed a method on DNA steganography using DNA insertion algorithm on paper "Highly Improved DNA Based Steganography". In this method, a binary key value is assumed and the secret message is split into several binary segments. First, perform XOR operation with the key and first segment and the result will XOR with next the segments and so on. Then the final value is inserted into randomly generate reference DNA sequence. The decryption process is the complement of encryption process.

III. PROPOSED FRAMEWORK

The four different DNA nitrogenous bases adenine (A), cytosine(C), guanine (G) and thymine (T) are assigned two-digit binary values uniquely using the DNA dictionary method. The bases A, C, G, and T are assigned the binary values 00, 01, 10 and 11 respectively. We convert the intended message into an 8-bit binary sequence according to their ASCII values and using DNA dictionary rule each 8-bit binary sequence is converted into a corresponding DNA sequence.

We select a random DNA sequence as the cover media. Then we construct a balanced tree where each node contains one character (base) of cover media. The size of random DNA is determined by the height of the designed tree where the height of the tree is determined by the size of intended message. [Table-I] shows the number of internal nodes and leaf node according to the height of the balanced tree.

Table-I: Number of nodes in the tree in terms of height

Height	Internal Nodes	Leaf Nodes	Total Nodes
1	0	1	1
2	1	2	3
3	3	6	9
4	9	12	21
5	21	36	57
6	57	72	129
7	129	216	345
8	345	432	777
9	777	1296	2073
10	2073	2592	4665

In our proposed method, we replace each leaf node value with the encrypted DNA message. The number of the leaf nodes of the tree will be equal to the size of encrypted DNA (intended message).

To make the tree structure unpredictable to the intruders, we generate a tree such that each internal node of the even level has at most two children, whereas each internal node of odd level has at most three children. The root of the tree is at level number zero. We will traverse the tree using Depth-First Search (DFS) also known as Euler tour in a tree [9] and number each node according to their discover time.

Finally, we get a fake DNA sequence (reference DNA) from the tree by ordering to their discover time. The decoding process is done by complementing the encoding process.

A. *The Steps of Encoding*

Step 1: Each character of the intended message is converted into an 8-bit binary string according to ASCII Value.

Step 2: Each binary string is converted into DNA sequence with the four DNA bases according to the DNA dictionary rule which generates an encrypted DNA sequence.

Step 3: A random DNA sequence is selected as cover media.

Step 4: We construct a balanced tree with the random DNA sequence.

Step 5: Replace all the leaf nodes of the tree with the encrypted DNA sequence.

Retrieval Number: D8088118419/2019©BEIESP
DOI:10.35940/ijrte.D8088.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Step 6: From the modified tree we get the reference DNA sequence.

The flow chart of our proposed method is shown in Fig. 1.

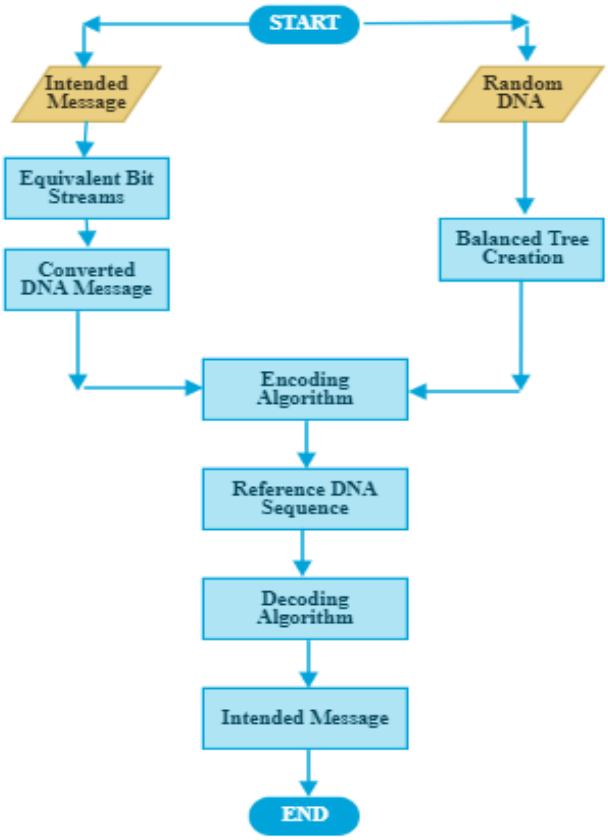

Fig. 1. Flow chart of the proposed algorithm

B. *An Example of Encoding Method*

Step 1: Consider the intended message "ABC" and convert the message into8-bit binary according to their ASCII value.

Table-II: Binary Representation of Secret Message

Character	ASCII Value	Binary Representation
A	65	0100001
B	66	0100010
C	67	0100011

Step 2: Convert each binary sequence into encrypted DNA message using the DNA dictionary rule [Table-III].

Table-III: Equivalent DNABases for each Character of Secret Message

Character	Binary Representation	Equivalent DNA Bases
A	01 00 00 01	CAAC
B	01 00 00 10	CAAG
C	01 00 00 11	CAAT

Thus, we get the encrypted DNA message of the intended message as "**CAACCAAGCAAT**" and the size of the encrypted DNA message is 12.

Step 3: A random DNA sequence is generated as cover media. Here, we select a random DNA sequence as "ACGGTTCCAATGCCTAAGCTA". Since the encrypted DNA message size is 12, so the corresponding random DNA sequence size will be 21 and the height of the tree will be 4[Level 0 – Level 3].

Step 4: We construct the balance tree with random DNA [Fig. 2, Table-IV].

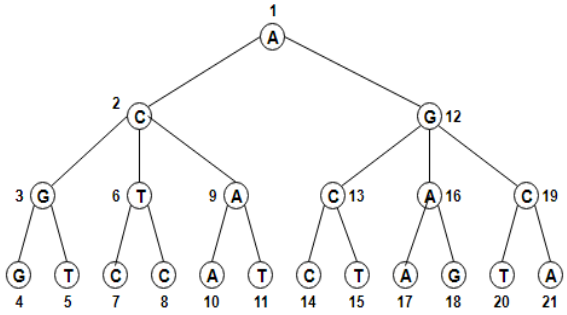

Fig. 2. Balanced Tree for random DNA (cover media) Sequence with Discovery Time

Fig. 3.

Table-IV: Discover Time of each Character of Random DNA sequence

Discover Time	Character	Discover Time	Character	Discover Time	Character
1	A	8	C	15	T
2	C	9	A	16	A
3	G	10	A	17	A
4	G	11	T	18	G
5	T	12	G	19	C
6	T	13	C	20	T
7	C	14	C	21	A

Step 5: Replace all leaf nodes of the sequence with the encrypted DNA message to get reference DNA sequence which is our secrete message [Fig. 3, Table-V].

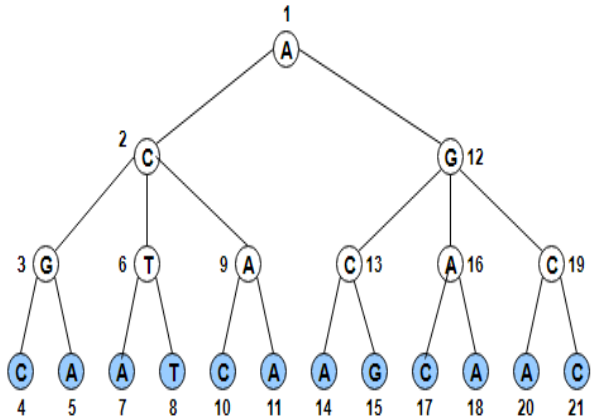

Fig. 4. Replaced Balanced Tree with Encrypted DNA Sequence (Secrete Message)

Table-V: Discover Time of each Character of Reference DNA Sequence

Discover Time	Character	Discover Time	Character	Discover Time	Character
1	A	8	T	15	G
2	C	9	A	16	A
3	G	10	C	17	C
4	C	11	A	18	A
5	A	12	G	19	C
6	T	13	C	20	A
7	A	14	A	21	C

Step 6: Finally, we get the reference DNA sequence from the modified balanced tree according to the discovery time in Depth-First Search. The reference DNA Sequence is:

'**ACGCATATACAGCAGACACAC**'

C. *Decoding Technique*

The receiver has to correctly manipulate the embedded DNA sequence to extract the intended message. The decoding process is the complement of the above encoding process. From the reference DNA sequence, we have to construct a similar balanced tree and extract the entire leaf nodes to get the encrypted DNA sequence. Using DNA dictionary rule conversion is done on the encrypted DNA sequence to get the 8-bit binary sequence. Finally, from the binary sequence, we will get the intended message.

D. *The Steps of Decoding*

Step 1: Construct a similar balanced tree with the fake reference DNA sequence.

Step 2: Extract the entire leaf nodes from the balanced tree to get the encrypted DNA sequence.

Step 3: Convert the encrypted DNA sequence to its binary value using the DNA dictionary rule and get a binary string.

Step 4: Extract each 8-bits binary sequence from left to right from the binary string.

Step 5: Convert each 8-bits binary sequence to its equivalent alphanumeric value according to ASCII standard.

Step 6: Finally, we get our intended message.

E. *An Example of Decoding Method*

Step 1: The fake reference DNA sequence is: "ACGCATATACAGCAGACACAC". Construct corresponding balanced tree with this received reference DNA sequence, [Fig. 4]. Fig-3 and Fig-4 must be the same.

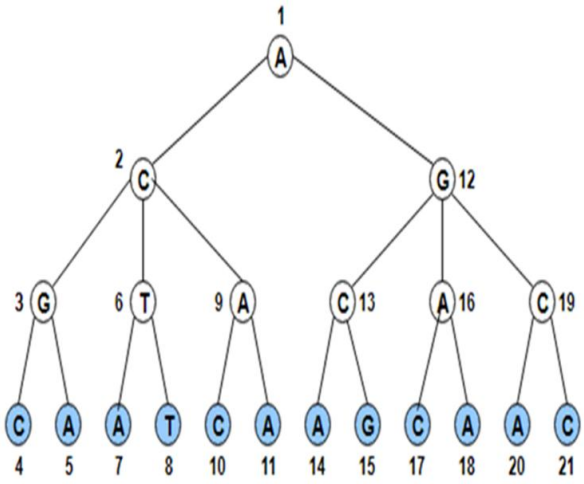

Fig. 5. Constructed Balanced Tree with received DNA Sequence

Step 2: Extract the entire leaf nodes from the balanced tree and get the encrypted DNA sequence "CAACCAAGCAAT".

Step 3: Convert the encrypted DNA sequence to its binary value using the DNA dictionary rule and get a binary string [Table-VI]. So, the equivalent binary bit streams are "010000010100001001000011".

Table-VI: Binary Value of Encrypted DNA Sequence

Character	Equivalent Binary Value
C	01
A	00
A	00
C	01
C	01
A	00
A	00
G	10
C	01
A	00
A	00
T	11

Step 4: Extract each 8-bits binary sequence from left to right from the binary string "010000010100001001000011".

01000001 **01000010** **01000011**

8-bits 8-bits 8-bits

Step 5: Convert each 8-bits binary sequence to its equivalent alphanumeric value according to ASCII standard [Table-VII].

Step 6: Finally get the intended message "**ABC**".

IV. EXPERIMENTAL RESULT

Our proposed method has been compared to other existing DNA based steganography techniques. The comparison distinguished with respect to embedded capacity known as payload capacity. Payload capacity is the measurement which is determined by how much intended message is being hidden within the cover media and usually measured in bpn (bit per nucleotide) [7].

Table-VII: Equivalent Alphanumeric Value According to ASCII Standard

Binary Bit Streams	Decimal Value	Equivalent alphanumeric value
1000001	65	A
1000010	66	B
1000011	67	C

$$\text{Payload Capacity (bpn)} = \frac{\text{Size of intended message}}{\text{size of cover media}}$$

We will always have a payload capacity of (P) ≥ 0.5.
In our proposed method,

At level 0, $LeafNode_0 = 1$ and $InternalNode_0 = 0$.
At level 1, $LeafNode_1 = 2$ and $InternalNode_1 = 1$.
At level 2, $LeafNode_2 = 6$ and $InternalNode_2 = 3$.
At level 3, $LeafNode_3 = 12$ and $InternalNode_3 = 9$.
The number of leaf nodes of this balanced tree can be expressed as a function.

$$LeafNode_i = \begin{cases} LeafNode_{i-1} \times 2 & , if\ i\ is\ odd \\ LeafNode_{i-1} \times 3 & , if\ i\ is\ even \end{cases}$$

$LeafNode_i = LeafNode_{i-1} \times K$, where $K \geq 2$

$InternalNode_i = InternalNode_{i-1} + LeafNode_{i-1}$

But it is guaranteed that, $LeafNode_{i-1} \geq InternalNode_{i-1}$

So, $LeafNode_i = LeafNode_{i-1} + LeafNode_{i-1}$ (if $K = 2$)
$\geq LeafNode_{i-1} + InternalNode_{i-1}$
$\geq InternalNode_i$

As, for K=2 it is proved that, $LeafNode_i \geq InternalNode_i$, It is also proved for K=3.So, for our proposed balanced tree data structures, $LeafNode_i \geq InternalNode_i$

Our proposed method in this paper provides higher payload capacity than other existing methods in DNA Steganography. The payload capacity of our methods is always ≥ 0.50 bpn. In our methods, the size of the reference DNA sequence is always less than two times of encrypted DNA sequence.

The methods suggested by Khalifa A [7], shows lower payload capacity than our proposed methods. The payload capacity of this method is at most 0.33 bpn. So, the reference DNA sequence must be three times larger than the secret DNA sequence. More precisely, if the size of the secret DNA sequence is n, then the size of the embedded reference DNA sequence must be at least 3n.

In the proposed method by Malathi P. et al. [8], the actual capacity is undefined. It depends on the key value K2. By higher key value of K2, the reference sequence becomes larger that trends to the low payload capacity. Approximately, the reference DNA sequence is K2 times larger than secret DNA sequence. If the randomly generated key value K2 is larger than three or four, then the size of a randomly generated reference DNA sequence will be extremely high for the long message. The following table [Table-VII] shows the comparison between our proposed method and other existing methods in terms of payload capacity.

Table-VIII: Comparisons Table in terms of Payload Capacity

Provider	Approach	Secret DNA Message Size	Reference DNA Size	Payload Capacity
Khalifa A [7]	LSBase	93312	279936	0.33
Malathi P. et al [8]	Insertion Method	93312	373248	0.25
Proposed Framework	Balanced Tree	93312	167961	0.56

Following Figure-5 shows that our proposed method outperforms against other methods in terms of payload capacity.

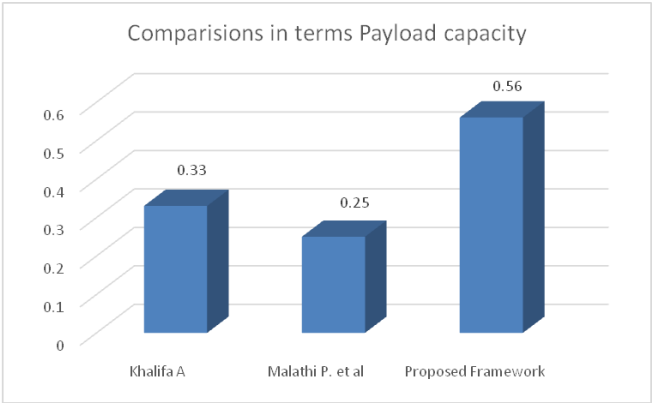

Fig. 6. Comparisons existing methods with proposed method

V. CONCLUSIONS

In this paper, we have introduced a new DNA steganography framework which shows superior performance in comparison with other existing techniques in terms of payload capacity. Where any existing method in DNA steganography does not have payload capacity more than (0.3-0.4), our proposed method shows payload capacity more than 0.50. The utilization of balanced tree property and hiding secret message in the leaf nodes ensures this outstanding payload performance in DNA steganography.

Future work concentrates on improving more payload capacity (>0.75) using data compression algorithms. Both cryptography and steganography techniques can be applied simultaneously to provide more payload, security, and robustness in data transmission.

REFERENCES

1. H. J. Shiu, K. L. Ng, J. F. Fang, R. C. T. Lee and C. H. Huang, "Data Hiding Methods Based upon DNA Sequences", Information Sciences, Vol. 180, No. 11, pp. 2196-2208, June, 2010.
2. S. Sun, "A Novel Edge Based Image Steganography with 2k Correction and Huffman Encoding", Information Processing Letters, Vol. 116, No. 2, pp. 93-99, February, 2016.

3. Malathi P, Gireeshkumar T., "Relating the Embedding Efficiency of LSB Steganography Techniques in Spatial and Transform Domains", Procedia Computer Science 2016; 93:878-85.
4. EihabBashier, Ghaidaa Ahmed, Hussam-aldeen Othman, Rayan Shappo, "Hiding Secret Messages using Artificial DNA Sequences Generated by Integer Chaotic Maps", International Journal of Computer Applications (0975 - 8887) Volume 70 - No. 15, May 2013.
5. Wang Z, Zhao X, Wang H, Cui G., "Information hiding based on DNA steganography", In 4th IEEE International Conference on Software Engineering and Service Science (ICSESS)2013;946-949.
6. S. Jiao, and R. Goutte, "Hiding data in DNA of living organisms" Natural Science, 1(3): 181-184, 2009.
7. Khalifa A. LSBase: "A key encapsulation scheme to improve hybrid crypto-systems using DNA steganography", In 8th IEEE International Conference on Computer Engineering & Systems (ICCES) 2013; 105-110.
8. Malathi P, Manoaj M, Manoj R, VaikunthRaghavan, Vinodhini R. E, "Highly Improved DNA Based Steganography",7th International Conference on Advances in Computing & Communications, ICACC-2017, Procedia Computer Science, Volume 115, 2017, Pages 651-659.
9. RE Tarjan, U Vishkin, "Finding biconnectedcomponemts and computing tree functions in logarithmic parallel time", in 25th Annual Symp. On Foundations of Computer Science, 1984, pp. 12-22.
10. "Applications of Intelligent Optimization in Biology and Medicine", Springer Nature America, Inc, 2016.
11. Haval I. Hussein, Wafaa M. Abdullah, "A Modified Table Lookup Substitution for Hiding Data in DNA", 2018 International Conference on Advanced Science and Engineering (ICOASE), 2018.

AUTHORS PROFILE

Partha Saha completed his B.Sc. (Engineering) in Computer Science and engineering from Mawlana Bhashani Science and Technology University, Bangladesh. His research interest is Algorithm Design and Analysis, Cyber Security, Machine Learning, and Artificial Intelligence.

Lubna Yasmin Pinky completed B.Sc. Hons and M.Sc. from Jahangirnagar University in Computer Science and Engineering and assistant professor at Mawlana Bhashani Science and Technology University, Bangladesh. Her fields of interest are Bio-Informatics, Artificial Intelligence and Deep Learning.

Mohammad Ashraful Islam completed B.Sc. Hons and M.Sc. from Jahangirnagar University in Computer Science and Engineering and lecturer at Jahangirnagar University, Bangladesh. His fields of interest are Algorithm Design and Analysis, Computer Vision, Robotic Vision, Image Processing and Deep Learning.

Papia Akter perusing her B.Sc. (Engineering) in Computer Science and engineering from Mawlana Bhashani Science and Technology University, Bangladesh. Her fields of interest are Algorithm Design and Analysis, Bio-Informatics, Cyber Security and Robotic Vision.