

# Data Hiding by Unsupervised Machine Learning Using Clustering K-mean Technique

Hiba Hamdi Hassan<sup>1</sup>, Maisa'a Abid Ali Khodher<sup>2</sup>

<sup>1,2</sup>Department of Computer Science, University of Technology, Baghdad, Iraq  
<sup>1</sup>cs.19.05@grad.uotechnology.com, <sup>2</sup>Maisaa.a.Khodher@uotechnology.edu.iq

**Abstract**— Steganography includes hiding text, image, or any sentient information inside another image, video, or audio. It aims to increase individuals' use of social media, the internet and web networks to securely transmit information between sender and receiver and an attacker will not be able to detect its information. The current article deals with steganography that can be used as machine learning method, it suggests a new method to hide data by using unsupervised machine learning (clustering k-mean algorithm). This system uses hidden data into the cover image, it consists of three steps: the first step divides the cover image into three clusterings that more contrast by using k-means cluster, the selects a text or image to be converted to binary by using ASCII code, the third step hides a binary message or binary image in the cover image by using sequential LSB method. After that, the system is implemented. The objective of the suggested system is obtained, using Unsupervised Machine Learning (K-mean technique) to securely send sensitive information without worrying about man-in-the-middle attack was proposed. Such a method is characterized by high security and capacity. Through evaluation, the system uses PSNR, MSE, Entropy, and Histogram to hide the secret message and secret image in the spatial domain in the cover image.

**Index Terms**— Steganography, (LSB), K-mean, Cluster, Machine learning.

## I. INTRODUCTION

One of the main methods you may secure your privacy is data concealing. The aim is to hide data by including sufficient personal data in maintenance. Many techniques are used in steganography. Spread spectrum, LSB, Transform ...etc. [1]. Steganography is, many times, confused with cryptography, yet they are different as steganography hides the data so that nothing appears out of the ordinary, while cryptography encrypts the text. Steganography must hide data in multimedia content such as files, texts, images, or videos. Steganography tries to conceal the fact that secret data is being transferred secretly as well as the contents of the secret data [2]. Machine learning is dependent on computer learning like human brain capabilities a computer-based learning method. Machine learning is supervised, unsupervised, and enhanced learning as three key categories [2][3]. The classifying approach is K-mean and is faster than hierarchical clustering, simple and computational. And work on many variables [1] [3].

In 2018, Zihan Wang, Neng Gao, and et.al, proposed a novel stenographic self-learning algorithm based on the adversarial generative network .its the method named SSteGAN. This approach uncontrollably explains the stenographic algorithm and generates a stego picture directly from the secret message without the cover image. Meanwhile, through comparison with comparable activities, it was evaluated the model performance and analyzed decryption security. In sum, the study results in anticipated results and opens up a new way of integrating generative opponent networks into steganography [2].

In 2021 AL HUSSIAN S. SAAD, M. S. MOHAMED, and et. al, proposed a coverless data hiding concept to solve this problem. Coverless does not indicate that the secret message is communicated or the cover file can be deleted without utilizing a cover file.

Received 1/6/2021; Accepted 20/8/2021

Instead, a cover file or secret message mapping is created to embed the secret message. This article proposes a result of a new, very rugged and extremely durable image steganography approach based on optical trademark identification (OMR) and machine rules-based learning (RBML). This method depends on the machine learning rule-based (RBML) and OR algorithms (OMR). The RBML algorithm was initially constructed and then students were simulated in the last T/F bladder examination, detected and marked based upon the co-circles [5].

The contribution that distinguishes the proposed system of this work is the use of (unsupervised machine learning (k-mean clustering algorithm) to hide data easily and quickly in performance, which increases the possibility of security of information inside images using methods of hiding easily and securely at the same time

## II. K-MEAN CLUSTER ALGORITHM

An algorithm for K-means Clustering is a way to split a set of data into certain categories. K-means clustering is one of the prominent methods. In k-means, a data collection is split into a group k number of data. It classifies a certain set of data into a distinct cluster k number. The K-means algorithm consists of two independent steps. In step one, the centroid k is calculated, and in step two, every point is taken to the cluster from the data point that has the closest centroid. There are various approaches to distinguish the nearest centroid, Euclidean distance is among the most common approaches. [1]

1. Initialize several clusters (k) and Centre (ck).
2. For each pixel of an image, calculate the Euclidean distance d, between the center (ck) and each pixel (p(x,y)) of an image using the relation given below.

$$d = \|p(x, y) - ck\| \quad (1)$$

3. Assign all the pixels to the nearest center based on distance-distanced pixels that have been assigned, recalculate the new position of the Centre (ck) using the relation given below.

$$ck = \frac{1}{k} \sum_{y \in ck} \sum_{x \in ck} p(x, y) \quad (2)$$

4. Repeat the process until it satisfies the tolerance or error value.
5. Reshape the cluster pixels into an image.

## III. STEGANOGRAPHY

The security of information and confidentiality of any kind of communication is the primary priority. Steganography is often known as way of using unusual digital media such as text, audio, video, and images to hide secret data. Steganographic system design is challenged by ensuring a fair deal between safety, robustness, a higher rate of incorporation, and imperceptibility. [14] The technique of steganography is to hide the letters that the attackers have not seen. The modifications are a letter, although it cannot be understood [17].

### A. Image Steganography

With digital pictures proliferating and a high level of redundancy (despite their compression) in the digital image representation, the interest has arisen in employing digital pictures as a cover object for steganographical purposes [18]. Every single pixel in the RGB color model has three red, green, and blue hues. The method PLEXING is the amplification of the colors [19].

### B. The Classical LSB Image Steganography Method

Steganography is an information hiding branch that hides a message inside a cover (the other branch is digital watermarking). It provides a form of subliminal channel for secrecy so as not to identify the presence of the message by the intended hacker or attacker. The LSB is a basic and simple way of data concealing. Several scientists then proposed the number of enhanced LSB algorithms [17]. LSB techniques in the picture steganography are complicated and secret keys control hidden information and cannot be retrieved without the same secret keys. It introduces a novel way to LSB picture steganography using hidden maps. Using secret keys to secure hidden information controls a Secret Map [20] [11].

## IV. EVALUATION SYSTEM PERFORMANCE

### A. MSE Mean square error (MSE)

MSE is one approach to quantifying the difference between the estimator values and the real quantity values. The following equation is used to compute MSE [8] [9]

$$MES = \frac{\sum [I_1(m,n) - I_2(m,n)]^2}{M * N} \quad (3)$$

Where m represents the number of rows and n is the number of columns inside the cover image.

### B. Peak Signal to Noise Ratio (PSNR)

PSNR is a typical measuring method to hide information for testing the quality of the stego images. The better the stego image quality, if the value of the PSNR is higher. The following equation is calculated for PSNR:

$$PSNR = 10 \frac{R^2}{MES} \quad (4)$$

R in the above equation is the greatest fluctuation of the type of input image data. For instance, if the image input has a floating-point type with double accuracy, R is 1 [7] [8][10].

### C. Capacity (CAP)

CAP That relates to the quantity of information in a picture. Capacity (message size) and solidity interact, and capability is one of the key quality parameters of steganography. [21].

$$\text{Ratio capacity} = \text{size of secret data} / \text{size of original media} \dots \quad (5)$$

#### D. Entropy

Entropy is a random measurement statistic that can be utilized to characterize the input image texture [12]. There is a need for a statistical metric for random analysis of picture block pixel values. Before randomization was determined, the pixel values in the image block were quantified to 64 levels. This quantification procedure contributes to the identical value of the extremely narrow pixels, which in turn helps to achieve a low entropy measurement from a genuine pixel block. The entropy definition is given in equations for a block of pixels 0 to 63 [15].

$$H(X) = -\sum_{i=0}^{63} P(x_i) \log_2 P(x_i) \quad (6)$$

This equation yields an estimate of the average minimum number of bits that is needed to encode a string of bits on the basis of the frequency of the symbol.

#### E. Histogram

The calculated image histogram is a figure that displays several pixels on the indexed image in each degree or indicator. To be able to achieve sensitive dissent [22]. The Histogramm incorporates data needed for the picture normalization when pixels are long. Histogram reveals each pixel's precise occurrence in the picture. The remarkable resemblance between the histograms of the host and the stego shows the minimum distortion after the secret picture has been integrated into the host image [13].

$$P(m, n) = \frac{\text{number of pixels with scale level} \leq (m, n)}{\text{Total number of pixel}} x(\text{maximum scale level}) \quad (7)$$

Where (m, n) represent the number of rows and the number of columns inside the cover image.

## V. THE PROPOSED FRAMEWORK

The proposed system consists of three steps. The flowchart of the proposed system, as shown in Fig. 1.(a, b), explains the embedding and extraction algorithms.

DOI: <https://doi.org/10.33103/uot.ijccce.21.4.4>

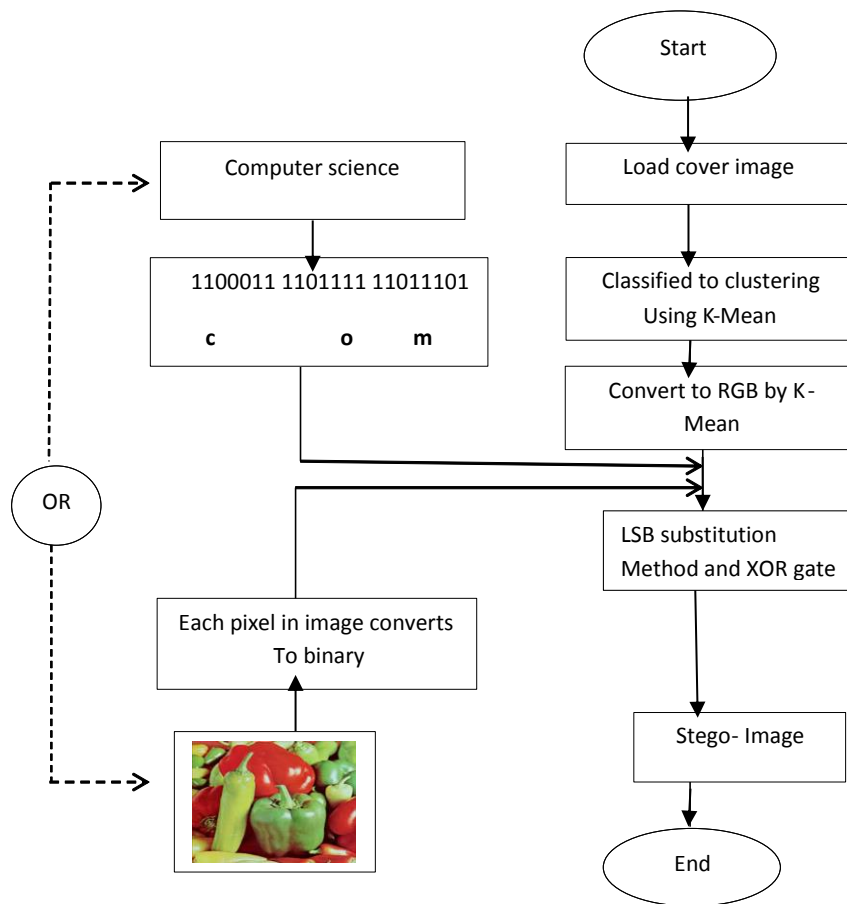


FIG. 1. THE FLOW CHART OF THE PROPOSED SYSTEM, A) EMBEDDED ALGORITHM

Received 1/6/2021; Accepted 20/8/2021

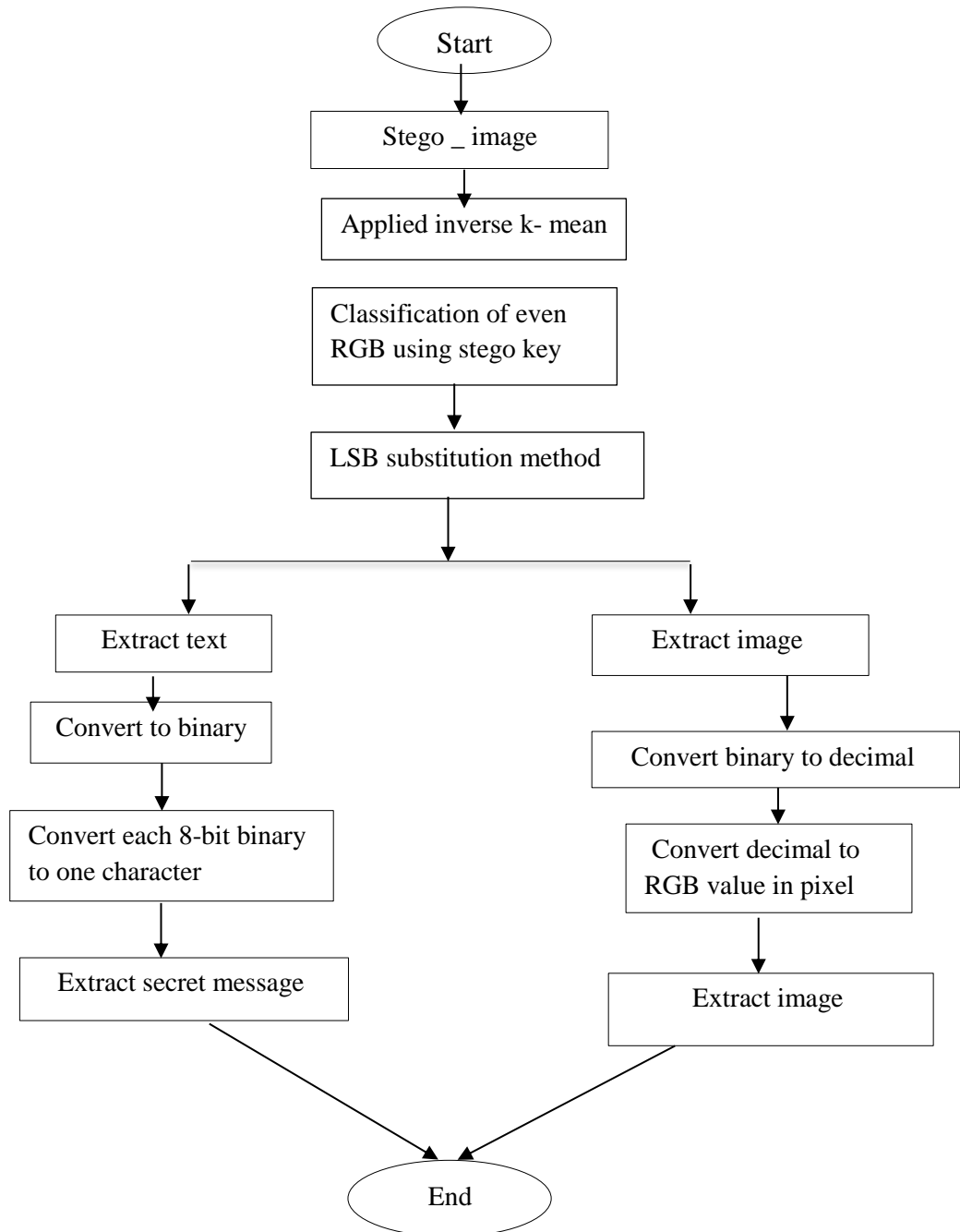


FIG. 1. THE FLOW CHART OF THE PROPOSED SYSTEM, B) EXTRACTION ALGORITHM

Received 1/6/2021; Accepted 20/8/2021

DOI: <https://doi.org/10.33103/uot.ijccce.21.4.4>**A. Embedded algorithm**

Input: cover image, secret image, secret message, K- mean.

Output: stego cover image.

Step 1: Read the cover image and the hiding text message and image.

Step 2: select best of color in the image and classify to three clusters by using k mean

Step 3: Convert text message or secret image into binary

Step 4: Determine the last bit of each pixel in the cluster of the cover image calculated LSB

Step 5: Replace LSB of cover image with each bit of secret (message or image) one by one

**B. Algorithm extract secret image or message**

Input: stego cover image.

Output: secret image, secret message.

Step 1: Read the stego image

Step 2: select best of color in the image and classify to thclustersster by using k mean

Step 3: Calculate LSB of each pixel of in clusters ego image

Step 4: Retrieve bits and convert each 8 bit into character

Where the steps of the propsed system

- **first step: Classifying an Image by using k means clustering**

This step is used image cover to hide data, this data is a secret message or secret image. The cover is classified by using machine learning unsupervised (Clustering k-means). To select the best color included in the image that contains more pixels within the image is done by converting a cover image to RGB pixel, the pixels are loaded into a data set and the image is opened. Every pixel value (r, g, b) is stored in an array. [(r1, g1, b1), (r2, g2, b2), (r3, g3, b3)..... (Rn, G n, Bn)] The pixels are split into clusters of K. A set of all elements, the tuples (R, G, B), used to determine each K cluster's centroid. The pixel set centers are determined by calculating the distance between each pixel pair and the central value is updated. Once the image center is identified, its values will be stored in several tuples (r, g, b)

$$D(k1, P2) = \sqrt{(r_1 - r_2)^2 + (g_1 - g_2)^2 + (b_1 - b_2)^2} \quad (8)$$

The pixels are separated into clusters, which are the center value closest to the pixel. In the arrays of each cluster, the pixels (r, g, and b) get attached. Then these arrays become images, as shown in Fig. 2. (a, b).

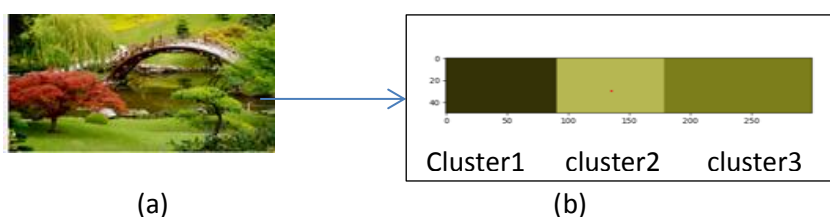


FIG. 2. THE CLASSIFIED COVER IMAGE, A) COVER IMAGE, B) RGB K-MEAN ALGORITHM.

Received 1/6/2021; Accepted 2018/2021

DOI: <https://doi.org/10.33103/uot.ijccce.21.4.4>

- **second step: select (secret message or secret image) and convert it to binary**

This step is to convert secret messages to binary by using ASCII. Each character represents 8-bit, and it is converted the secret image to direct binary, each pixel in the image is a decimal number, convert each number to a binary represented to 8-bit, as shown in Fig. 3. (a, b). For Example

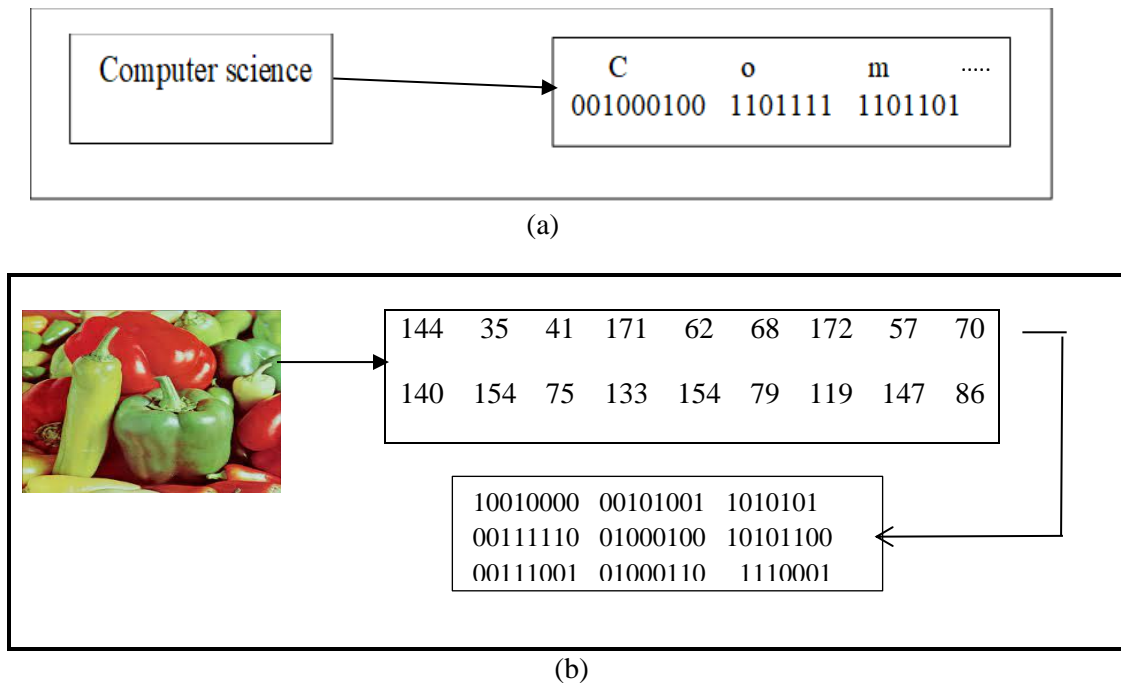


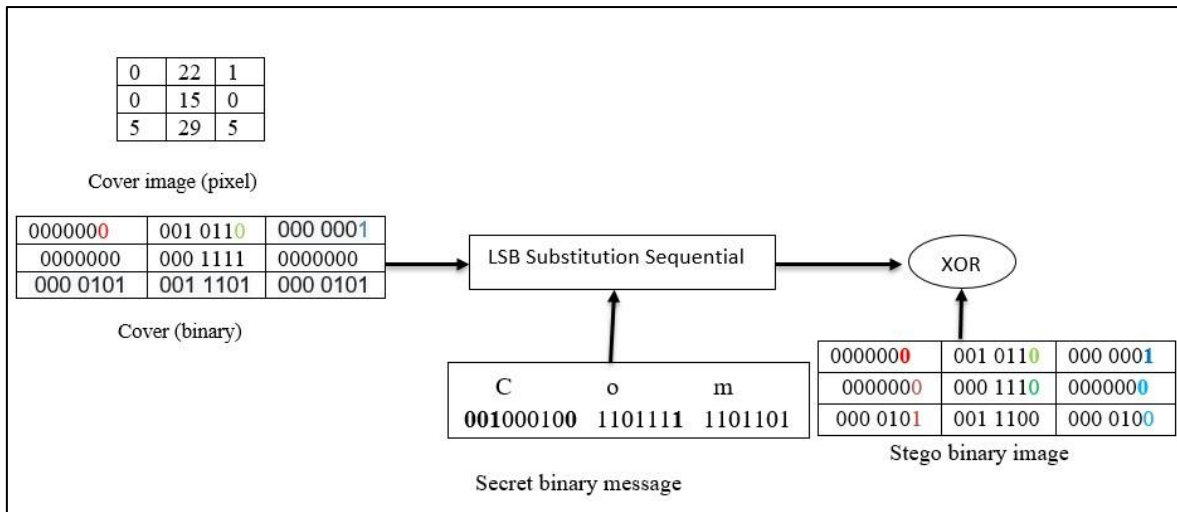
FIG. 3. CONVERT DATA TO BINARY A) CONVERT SECRET MESSAGE TO A BINARY IMAGE, B) CONVERT SECRET IMAGE TO BINARY

- **Third step: hide a binary message or binary image in the cover image by using LSB the method**

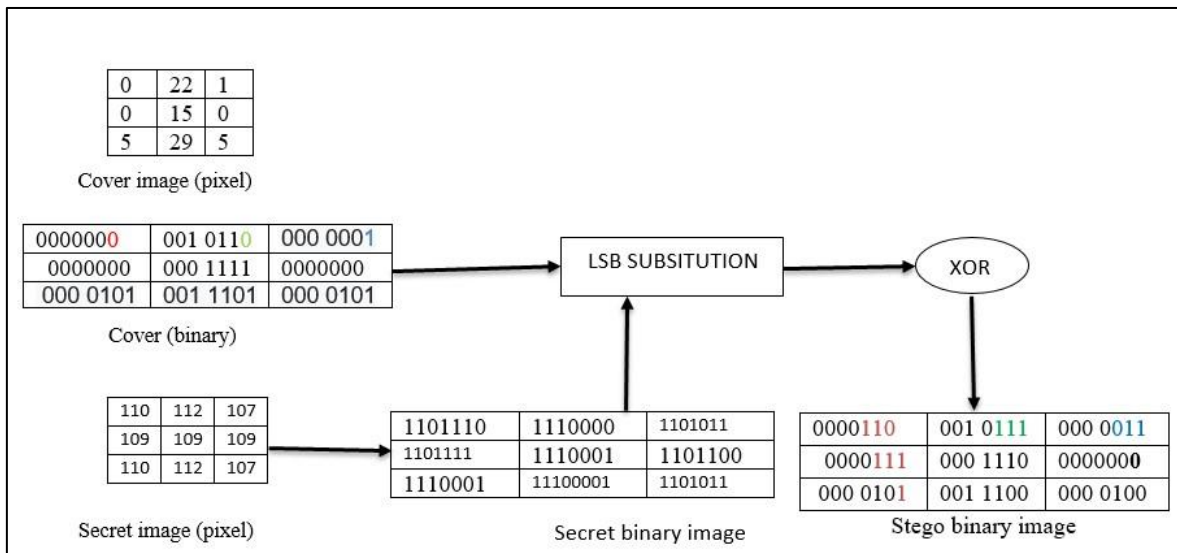
This stage involves selecting a message or image to be hidden in a cover based on the size of the cover and then hiding it using the LSB technique (The LSB technique directly hides the secret binary sequence in the cover image, which sequentially replaces numbers of the last bit into the cover image with the secret bit numbers to hide each bit in cover.) For each RGB component, there is a need to perform an operation called "XOR" that combines a sequential secret message bit with the 7th bit, then place the result of this operation in the last bit of the cover picture to create the stego object.

Received 1/6/2021; Accepted 2018/2021





(a)

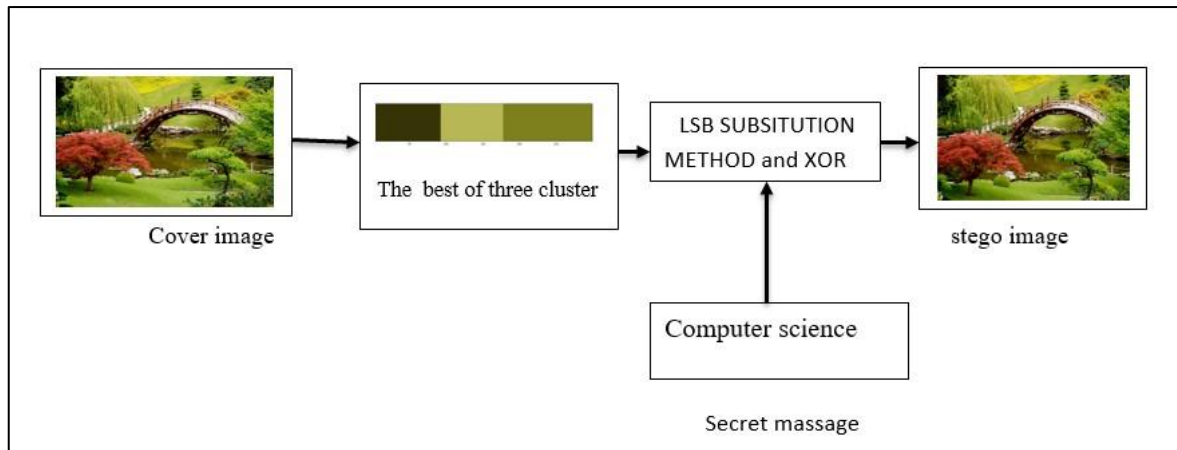


(b)

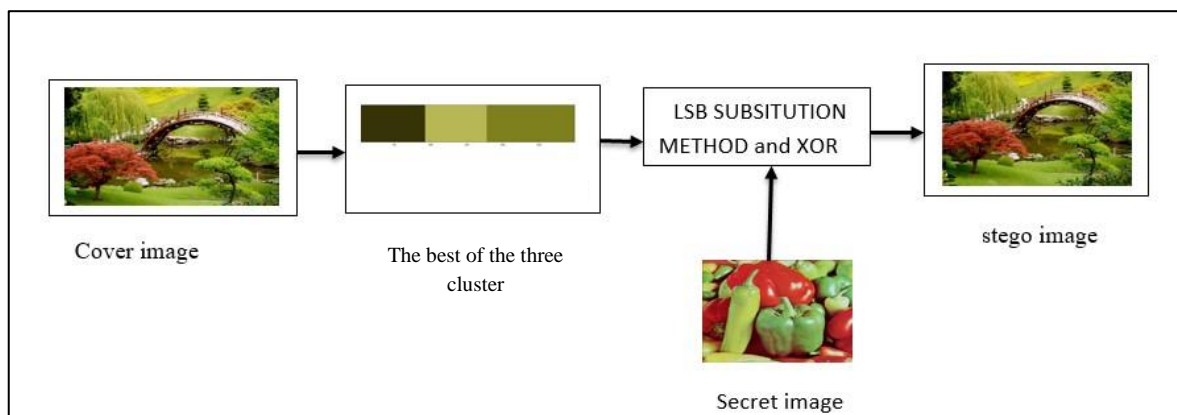
FIG. 4. (A) THE BINARY SECRET MESSAGE HIDES EACH BIT 0 OR 1 IN THE LAST BIT IN LSB IN THE LAST PIXEL OF COVER USING XOR GATE, (B) AND THE BINARY SECRET IMAGE IS HIDING EACH THREE-BIT FOR 0 OR 1, IN COVER IMAGE IN LSB IN EACH PIXEL USING XOR GATE

## VI. IMPLEMENT THE SYSTEM

Fig. 5 shows implementing the system data hiding by method (unsupervised K-mean clustering). explains implementing the system hiding data hiding for the text and the image to obtained stego image.



(a)



(b)

FIG. 5. (A) IMPLEMENT SYSTEM FOR A SECRET MESSAGE, (B) IMPLEMENT SYSTEM FOR A SECRET IMAGE

## VII. TESTING THE RESULTS

This section discusses the outcomes of the suggested system for using a set of tests, which are MSE, PSNR, entropy, RMSE, histogram, and time. Table 1 shows the implementation of the suggested system for hiding a secret message or secret image, as the second column shows the size of the message or the image to be hidden. The third column shows the cover image, and the fourth column shows the colors cluster that appear the most in the image by using K-mean algorithm, and these colors are the ones that will be hidden, and the last column shows the shape of the cover image after hiding an image inside. It shows us that the image has no distortion and has kept the same size. Table 2 indicates the measurements for hiding a secret message in the cover, Table 3 indicates the measurements for hiding a secret image in the cover. That system is good in hiding messages or images in the cover through testing the cover that was embedded. Table 4 indicates different histograms for a cover embedded message or image.

- The analysis of the suggested system is done to obtain results for a secret message, the range of MSE is from  $7.006e-14$  to  $7.604e-13$ , the range of PSNR extends from 79.8600 to 80.2063, the range of time extends from 8.11403 to 7.58068, and the range of capacity extends from 0.0262 to 0.0260. The MSE is decreased and the PSNR increased.

Received 1/6/2021; Accepted 20/8/2021

DOI: <https://doi.org/10.33103/uot.ijccce.21.4.4>

• The analysis of the suggested system is to obtain results for the secret image, the range of MSE is from 0.0636 to 0.0676, the range of PSNR from 29.3015 to 29.2506, the range of time from 9.8518 to 9.85289, the range of capacity from 0.02704 to 0.0461. The MSE is decreased and the PSNR is increased.

TABLE 1. RESULT IMPLEMENTATION OF THE PROPOSED SYSTEM FOR HIDING A SECRET MESSAGE OR SECRET IMAGES AND CLUSTER K-MEAN.


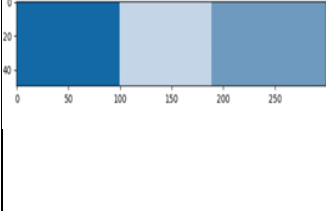


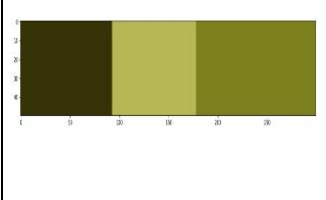


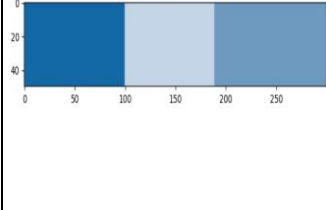


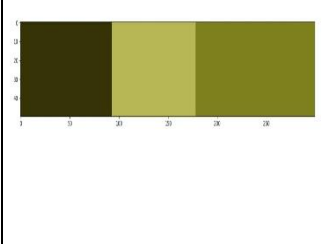

No. of message or image	secret message or image	Original cover image	Stego k-mean cover	Stego cover image
1	Secret message 170 bytes			
2	Secret message 198 bytes			
3	Secret image 150 x 150			
4	Secret image 225 x 225			

TABLE 2. INDICATED THE MEASUREMENTS FOR MSE, PSNR, ENTROPY, CAPACITY, AND TIME, TO HIDE THE SECRET MESSAGE.

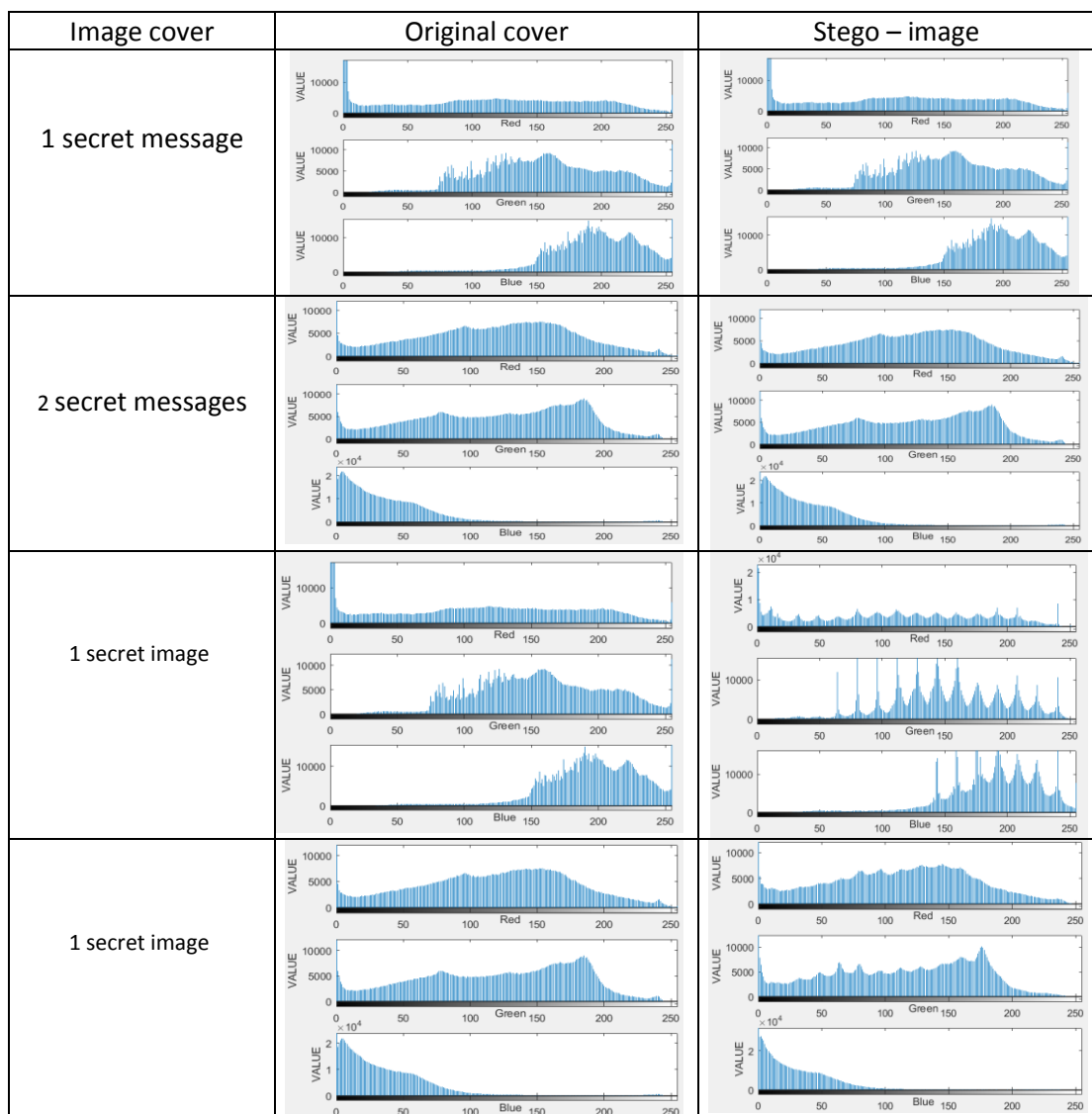
Image cover	MSE	PSNR	ENTROPY	Capacity	TIME
1	7.006e-14	79.8600	7.689508	0.0262	8.11403
2	1.755e-13	82.6153	7.6824	0.0277	23.369
3	7.604e-13	80.2063	7.3693	0.0260	7.58068

TABLE 3. INDICATED THE MSE, PSNR, ENTROPY, CAPACITY, AND TIME, FOR HIDING SECRET IMAGES.

Image cover	MSE	PSNR	ENTROPY	Capacity	TIME
1	0.0636	29.3015	4.0099	0.02704	9.8518
2	0.0428	29.648	4.0579	0.0570	16.4324
3	0.0676	29.2506	4.057	0.0461	9.85289

Received 1/6/2021; Accepted 20/8/2021

TABLE 4. INDICATES DIFFERENT HISTOGRAMS FOR THE COVER EMBEDDED SECRET MESSAGE OR SECRET IMAGE.



### VIII. CONCLUSIONS

This paper uses an Unsupervised machine learning method (k-mean clusters) to hide the data (message or image). This algorithm divides the image cover into three clusters of RGB depending on the colors to have more density in the image and selects a location to hide one bit from the secret message or three-bit from a secret image by using the spatial domain sequential LSB algorithm. The outcomes of the suggested system obtain better images from text, because of high quality, efficiency, robustness, and high security. Through a set of tests PSNR, MSE, entropy, and robustness measure the histogram. In the future to have better results, more complex steganography and encryption techniques can be used to improve the work

Received 1/6/2021; Accepted 20/8/2021

## REFERENCES

- [1] Dhanachandra, Nameirakpam, Khumanthem Manglem, and Yambem Jina Chanu. "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm." *Procedia Computer Science* 54 (2015): 764-771.
- [2] Jung, K. H. A study on machine learning for steganalysis. In *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing* (2019, January). (pp. 12-15).
- [3] Yang, Chyuan-Huei Thomas, and Wen-Feng Wu. "Data hiding method in color image based on grouping palette index by particle swarm optimization with k-means clustering." *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldCom), 2011*.pp 1-6.
- [4] Kim, Jaeyoung, et al. "Patent document clustering with deep embeddings." *Scientometrics* s (2020) 123:563–577.
- [5] Sasmal, M. M., & Mula, M. D. An Enhanced Method for Information Hiding Using LSB Steganography. In *Journal of Physics: Conference Series* (2021, February). (Vol. 1797, No. 1, p. 012015). IOP Publishing, Pp1-8.
- [6] Jiang, Nan, Na Zhao, and Luo Wang. "LSB based quantum image steganography algorithm." *International Journal of Theoretical Physics* 55.1 (2016), pp 107-123.
- [7] Namath, Irtefaa A., and Hind Rustum Mohammed. "Text Steganography in Statistically Clustered Iris Image." *EAI Endorsed Transactions on Energy Web Online* First.18-11-2020.167100, pp 1-7.
- [8] Sender, Levent, and Ivan W. Selesnick. "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency." *IEEE Transactions on signal processing* 50.11 (2002), pp 2744-2756.
- [9] Mahdi, Bashar S., and Alia K. Abdul Hassan. "Hybrid Techniques for Proposed Intelligent Digital Image Watermarking." *Eng. &Tech. Journal*, Vol.33, Part (B), No.4, 2015, pp702-713.
- [10] Kareem, Abdulameer A., and Abdul Monem S. Rahma. "A Statistical Image Noise Removal Adaptive Filter Using Rejection Test with F-Distribution." *Eng. &Tech. Journal*, Vol. 32, Part (B), No.2, 2014, pp302-312.
- [11] Mohsin H. AL-Zohair "Information Hiding image-Based on Random Locations" *Engineer at Ministry of Higher Education and Scientific Research, Baghdad, Iraq IJCCCE*, Vol.12, No.2, 2012, PP(9-15)
- [12] ALabaichi, Ashwak, Maisa'A. Abid Ali K. Al-Dabbas, and Adnan Salih. "Image steganography using least significant bit and secret map techniques." *International Journal of Electrical and Computer Engineering (IJECE)* Vol. 10, No. 1, February 2020, pp 935-946
- [13] Zebari, Dilovan Asaad, et al. "Image steganography based on swarm intelligence algorithms: A survey." *The Mattingley Publishing Co., Inc. May-June 2020 ISSN: 0193-4120 Page No. 22257–22269.*
- [14] Manikandan, V. M., and V. Masilamani. "Reversible data hiding scheme during encryption using machine learning." *Procedia Computer Science* 133 (2018), pp 348–356.
- [15] Honasan, H. S., et al. "A review of artificial intelligence techniques in image steganography domain." *Journal of Engineering Science and Technology Special Issue on ISSC'2016*, April (2017), pp103 – 113.
- [16] Masa's Abid Ali, K. Al, Ashwak Alabaichi, and Ahmed Saleem Abbas. "Dual method cryptography image by two force secure and steganography secret message in IoT." *TELKOMNIKA* 18.6 (2020), pp2928-2938.
- [17] Sencar, Husrev T., Mehdi Kharrazi, and Nasri Memon. "Image Steganography: Concepts and Practice." *Department of Electrical*. pp 1-31.
- [18] Abdulwahab, Hala Bahjat, Khaldoun L. Hameed, and Nawaf Hazim Barnouti. "Video Authentication using PLEXUS Method." *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS* 9.11 (2018).pp 730-737.
- [19] ALabaichi, Ashwak, Maisa'A. Abid Ali K. Al-Dabbas, and Adnan Salih. "Image steganography using least significant bit and secret map techniques." *International journal of electrical & computer engineering* 10.1 (2020), pp (2088-8708)
- [20] Msallam, Mohammed Majid. "A Development of Least Significant Bit Steganography Technique." *IRAQI JOURNAL OF COMPUTERS, COMMUNICATIONS, CONTROL, AND SYSTEMS ENGINEERING* 20.1 (2020)PP( 31-39).
- [21] Sehgal, Nancy, and Ajay Goel. "Evolution in image steganography." *International Journal of Information & Computation Technology* 4 (2014): 1221-1227.