FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 32, N^{o} 1, March 2019, pp. 147-177
https://doi.org/10.2298/FUEE1901147R

A NOVEL PRIORITY BASED DOCUMENT IMAGE ENCRYPTION WITH MIXED CHAOTIC SYSTEMS USING MACHINE LEARNING APPROACH

Revanna C R[1], Keshavamurthy C[2]

[1]Jain University, Bangalore, and Faculty of ECE,
Government Engineering College, Ramanagara, Karnataka, India
[2]Faculty of ECE, Sri Revanasiddeshwara Institute of Technology, Bangalore, Karnataka, India

Abstract. *Document images containing different types of information are required to be encrypted with different levels of security. In this paper, the image classification is carried out based on the feature extraction, for color images. The K-Nearest Neighbor (K-NN) method of image classification technique is used for classifying the query Document with trained set of features obtained from the Document database. Optical Character Recognition (OCR) technique is used to check for the presence as well as location of text/numerals in the Documents and to identify the Document type. Priority level is assigned in accordance with the Document type. Document images with different priorities are encrypted with different multi-dimensional chaotic maps. The Documents with different priority levels are diffused with different techniques. Document with highest priority are encrypted with highest level of security but Documents with lower priority levels are encrypted with lesser security levels. The proposed work was experimented for different document types with more number of image features for a large trained database. The results reveals a high speed of encryption for a set of document pages with priorities is more effective in comparison with a uniform method of encryption for all document types. The National Institute of Standards and Technology (NIST) statistical tests are also conducted to check for the randomness of the sequence and achieved good randomness. The proposed work also ensures security against the various statistical and differential attacks.*

Key words: *Ikeda, Lorenz, Chaotic, Feature Space, NIST*

Received August 4, 2018; received in revised form December 18, 2018

Corresponding author: Revanna C R
Research Scholar at Jain University, Bangalore, and Faculty of ECE, Government Engineering College, Ramanagara, 562159 Karnataka, India
(e-mail: revannacr2008@gmail.com)

1. INTRODUCTION

A document is any piece of information in text, picture or both forms on any medium which serves as a proof of evidence of a fact which may be secret, private or public. The advancement in digital and web based technology has led to document imaging (converting the physical form to electronic form) occupying prime importance and wide acceptance as the most reliable form of preserving data. In document management systems, one of the most important service is the sharing of documents in its image form which enables smooth functioning of an organization, where the documents are to be stored, retrieved, integrated, authenticated, distributed, collaborated, searched, reproduced, and transferred between members of the team. Due to the advancement in the internet technology, the distribution of documents containing confidential information over open network / wired or wire-less communication channels, offers wide scope for the interceptor to attack and hack the information. To overcome this problem there is a need to transform the intelligent information to an unintelligent form and distribute it over the channel. In all such cases the confidentiality, integrity, authenticity and non-repudiation of information is required to be maintained. The cryptographic technique called encryption, which transforms the intelligent form of information to unintelligent form is used to provide these kind of security during document sharing. Encryption is a ciphering technique which uses confusion and diffusion processes. The classical method of encryption is not appropriate as the document images are different from the electronic documents by their characteristics such as highly voluminous data, a strong correlation between neighboring pixels and redundant nature. The traditional lightweight block ciphers such as Advanced Encryption Standard (AES) and RC5 encrypts 128-bit blocks, the Data Encryption Standard (DES), the Triple DES and the Blowfish methods encrypts only 64-bit blocks, the MD4 and MD5 encrypts 512-bit blocks. These lightweight block ciphers are not appropriate to encrypt document images with larger block sizes. The chaotic systems used for encryption and decryption enable to use variable block size depending on the requirements as these images are voluminous, highly correlated and more redundant. Document images are classified according to user defined classes. Encryption of document images of different classes using a common single encryption algorithm vary in system resources (such as encryption time, complexity in design, computation complexity etc.) and utilization. Basically, providing security for all images is not required, but only the document images on demand are required to be encrypted. Images are different from each other by their features. Rather than encrypting the Document images of different types using a single method, classify them into different predefined types and encrypt those using different methodologies. This may take lesser encryption time, provide variable security level for each document type and make it difficult for the crypt analyzer to extract the key from the known cipher text-plain text attack. To encrypt and decrypt document images, the confusion and diffusion system is used with chaotic systems. The chaotic system is a non-linear dynamical system which generates pseudorandom numbers/sequences based on the initial conditions and system parameters. The chaotic sequences are aperiodic, random, deterministic, and sensitive to initial conditions and system parameters. The encryption of documents with different multi-dimensional chaotic maps may result in added security and increase the speed of encryption.

The proposed work is for classification and assignment of a document to one of the predefined set of classes and encrypt each class of documents with a different level of security to conserve the system resources. It uses the machine learning approach to classify the documents with an efficient and simple method of classifier with image features along with the OCR technique to differentiate text in them. The encryption is performed with confusion and diffusion process by using various multi-dimensional chaotic maps with different methodologies. Document image classification enables a fast retrieval of image document to encrypt when a large set of heterogeneous document images are present in the database. The various predefined set of document image classes considered are 1. The complete text document such as business letter, newsletter etc. 2. A complete graphical picture document such as photo, engineering drawing, pictures, diagrams etc. 3. A document with text embedded over the picture such as certificates 4. Medical image document showing the images of MRI, X-ray etc. 5. Text labelled images 6. Numerical value labelled images 7. Picture images with captions 8. Picture images with descriptive text such as newspaper 9. Landscape images 10. Animal class of images 11. Nature or scenic images 12. Images of buildings etc. The documents are classified based on the constitution of the document classes, the options available in document features and the chosen classifier algorithm. There are different methods to classify the document images. The K-NN method of classifier is chosen for classification. Based on the image features such as entropy, mean, variance, mean square error, skew, correlation, histogram, OCR and energy of the documents, the K-NN classifier assigns priority values. The images with least randomness are assigned the highest priority but the images with least OCR are assigned with lowest priority. Images with different priorities are subjected to different encryption methodologies. Each method of encryption uses confusion and different diffusion technique with different dimensional chaotic maps. The images with highest priority are encrypted with highest level of security when compared to lower priority images. The images with lower priorities are subjected to algorithms which can reduce the complexity and the encryption time.

2. LITERATURE SURVEY

Recognizing of documents depending upon their characteristics features are very important in Document management systems. A document analysis and recognition (DAR) [1] consists of different processing steps, Layout analysis, Character recognition, analysis of structural images containing textual information and its applications can be used for efficient document mining technique.

Content based image retrieval (CBIR) search technique make use of characteristic features that can be deployed for query document retrieval. A medical image retrieval system [2] using CBIR is performed by extracting visual features.

A study on document image classification is very important wherein the choice of different document image classifiers is based on the problem, the use of training data, the choice of document features [3]. The document in this context relates to a single page type-set document including a broader classification based on variations like business letters, articles and printed newspapers. The role of document classifier in document retrieval system is very important. The strength of the document classifier is based on image-level features, structural of textual features.

Document indexing in industrial context is very important where large number of documents are digitized every day and clustered in different classes. The document cluster can be achieved using a clustering technique. K-means. A well-known clustering technique [4] used for document clustering. The clustered data are classified using Content based image retrieval (CBIR) search technique which is based on the feedback learning.

Genre identification [5] in documents such as technical papers, photos, slides and tables is also important in document recognition system. The Document Genre identification be achieved by using machine learning approach by combining text based features and SVM.

Color Document images can be classified using color histogram features [6]. A large image database containing pictures of landscape, buildings, animal class, etc. can be classified using this technique. The classification of images is based on the color histogram features of images using the K-NN method of machine learning classifier results in an accuracy of 85%.

Textual information should be separated from non-text areas in the Document images. A block based Segmentation technique [7] is introduced to separate these text and non-text regions. Further Optical Character Recognition (OCR) system is deployed to identify the density of text from image pictures.

Comparison of different pages in a Document called page similarity is also important in Document analysis and classification technique. Page similarity can be achieved by using visual saliency metric defined on the basis textual parameters [8]. The obtained parameters are used to classify the Documents using K-nearest Neighbor classifier (KNN) [9].

OCR and machine learning approach such as Decision Tree [10] are combined to classify and for indexing heterogeneous document images. Color co-occurrence Matrix (CCM) [11] can be calculated for efficient image retrieval system. The Hue Saturation Value (HSV) is used for each pixel value of the image and the CCM is calculated by using the relevant formulae. The CCM of the sample image is compared with the images in the database and the resulting images are sorted based on the similarity. This method has the advantage of increased retrieval accuracy as the documents are retrieved based on the pixel information and color feature.

There are many document images which have the text or the numerals embedded over the picture. In such cases the presence of the text/numeral value is recognized by using the OCR (Optical Character Recognition) technique. The algorithm [12] Discussed extract the lines and curves which the alphabet is made is using the feature recognition based OCR. In [13] four different OCR techniques namely vectors crossing, Zoning, combination of vector crossing and Zoning and Template matching are proposed. The results obtained from these four methods are compared and it is shown that template matching technique yields better accuracy of 99.5% with an average time of 1.95 m sec per each character.

In this proposed work, once the query/sample document image is classified, it is subjected to appropriate encryption algorithm. There are different encryption techniques. The confusion and diffusion processes involved in encryption makes use of random numbers. The confusion process scrambles the image pixels and the diffusion process create the interdependency among the pixels. The random number generators are mainly the chaotic maps. A mixed chaotic system uses two different chaotic maps one for confusion and other for diffusion in an encryption algorithm.

The document image [14] Proposed is provided with a security using a mixed chaotic system in which the Document image is confused using a 1D-Logistic map and diffusion using 2D-Henon map against both the statistical and dynamical threats. A selective image encryption [15] Proposed yields with an NPCR=100% and UACI=33% which is close to the ideal values is achieved by using different diffusion techniques in a mixed chaotic system. A 2D-Ikeda chaotic map and 1D-Quadratic map are used for Confusion and Diffusion respectively. In [16] qualitatively estimated the complexity of random sequences statistically generated by the different chaotic maps. The random number sequences generated for the proposed system using different chaotic maps are tested for their randomness using NIST test for all the sixteen parameters [17].

From the literature, it is found that there is a need for a document management systems to encrypt document images of different types with, different levels of security, reduce the total encryption time and make it difficult for the cryptanalysis. In almost all the systems of image encryption, to the best of my knowledge, the same encryption technique is used irrespective of the document type. Here we are proposing a novel method, for reducing the total encryption time, providing varying security levels and tough cryptanalysis for a set of different types of document images. The proposed system is aimed at classifying the document images containing picture, text, text as caption, text embedded over picture, etc. and assign a priority value based on the type of information contained in them. The statistical features extracted for the document images are used for classification using K-NN image classifier with machine learning approach. The classified documents are assigned with a pre-defined priority value and subjected for encryption. The priority value of the classified document determines the method of encryption. Each priority is associated with different encryption methodology to obtain ciphered images with different security levels using different multi-dimensional chaotic maps. The chaotic maps used during the encryption and decryption are same.

3. Process Flow

Classification of document, depending upon their characteristics features are very important in Document management systems. In a large document database system, the document can be classified as, document containing only text, document contains only picture, document contains text embed over picture, labelled document, document contains picture with more text, document containing picture with caption, medical image documents, satellite image documents, Natural image document sets etc. In order to classify these documents, Document mining has to be performed. Typically Document mining involves mainly two process, namely the feature extraction and the feature Classification. Feature extraction is a very important step in pattern (Document Pattern) classification as it involves extraction of important feature vector which can be used to categorize one class of document with other class. The obtained feature vectors are classified using a classifier. The classifier compares the query document features with trained features and results in the class name into which the query document belongs to. The process flow diagram is shown in figure 1.

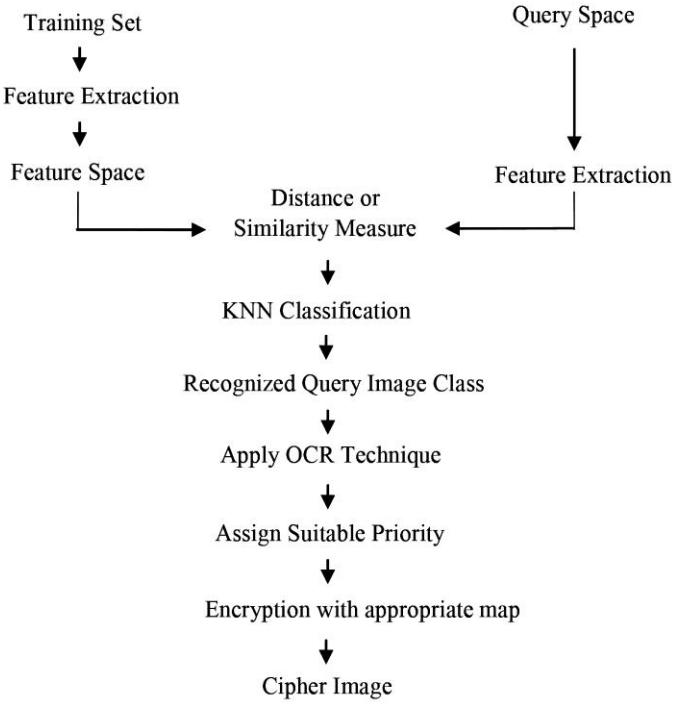


Fig. 1 The process flow diagram.

4. Feature Extraction

A color histogram features of an image are calculated in order to classify different documents. The extraction of color histogram of an image has an advantage over other technique since it requires less computation time as well as more efficient than other techniques.

Histogram Features

An image's histogram is a graphical representation plotted between number of grey levels (0 to 255) and pixel quantity at each grey level. The histogram graph depicts the nature of the image. An Image with uniform/flat histogram indicates that the image contains non correlated pixels (Encrypted image). Image with non-uniform histogram indicates that the image contains correlated pixels (Plain image). In a correlated image if the histogram is too sharp it indicates an image having low contrast. In the similar plain image if the histogram spreads over entire grey levels with non-uniform peaks indicates that it is an image having high contrast. Histogram features are considered as the statistical features, where the probability distribution of different intensity levels are modeled using histogram. These statistical features provides different traits pertaining to how the level of image intensity is spread. The histogram probability can be defined as

$$P(l) = \frac{N(l)}{M} \tag{1}$$

Where M, represents the total number of pixels in the image and $N(l)$ is the sum total of pixels at each grey level l. The value of histogram probability $P(l)$ lies in the range from 0 to 1. The sum of all probability $P(l)$ is equal to 1. Different features are considered depending on the probability of pixels in the histogram, Mean, Variance, Standard Deviation, Energy, Entropy and Skew Symmetric.

Mean

The mean indicates the average value calculated for all the pixels in an image, hence it provides brightness of the image. Image with high mean value indicates more brightness, whereas image with less mean indicates less brightness. The mean of an image for K=256 intensity levels can be calculated as

$$\bar{l} = \sum_{l=0}^{K-1} l \times P(l) = \sum\sum \frac{I(r,c)}{M} \tag{2}$$

Variance

Variance is a measure of contrast in an image. The region which has high contrast indicates high variance, whereas the region which has low contrast indicates low variance in an image. The variance can be calculated as

$$VAR = \sum_{l=0}^{K-1} (l - \bar{l})^2 \times P(l) \tag{3}$$

Standard Deviation (SD)

Standard Deviation can be obtained by applying the square root for the variance. It can be mathematically expressed as

$$\sigma = \mathrm{STD} = \sqrt{VAR} \tag{4}$$

Skew

Skew measures the asymmetry of probability distribution of pixels about its mean value. The value of the skew can be positive or negative or undefined. The value of skew is positive when histogram is spread to the right, and negative in case it is left tailed. Mathematically it can be expressed as

$$SKEW = \frac{1}{\sigma^3} \times \sum_{l=0}^{K-1} (l - \bar{l})^3 \times P(l) \tag{5}$$

Energy

Energy measures the distribution of intensity levels in an image. The value for Energy lies between 0 and 1. For an image the energy value is equal to 1 if it contains constant pixel value, and gets decremented if the pixels values are distributed among different intensity levels. Typically, for an image having more energy its compression ratio is high. Mathematically the energy can be expressed as

$$ENERGY = \sum_{l=0}^{K-1} P(l)^2 \tag{6}$$

Entropy

Information entropy is a measure of randomness in the image. The randomness of the image is based on the probability of occurrence of the various gray levels in the image. An image with all pixels of equal gray levels are equally probable represents the highest entropy. Mathematically the entropy can be expressed as

$$ENTROPY = - \sum_{l=0}^{K-1} P(l) \text{ x } log_2[P(l)] \tag{7}$$

Correlation

The images are characterized by high redundancy and significant correlation between adjacent pixels. Correlation mainly find the similarities between textures of two images or within an image. Hence used to find the image redundancies. To do this, the horizontal, vertical and diagonal correlation coefficients are required to be calculated. The smaller value of correlation coefficient between adjacent pixels represents the image is non-correlated and contains more random pixels. For a plain image, the correlation of adjacent pixels is nearly equal to one but for a text image, the correlation coefficient is less.

5. Feature Vectors and Feature Space

In a machine learning and pattern recognition technique, an image can be represented as n-dimensional symbolic or numerical values called feature vectors, where n represents the feature quantity. Feature measurements may either be numerical or symbolic in nature or sometimes both. A case in study for numerical feature is, calculating afore mentioned statistical values such as Mean, Standard Deviation, Variance, Energy, Entropy etc. for an image and storing these values in a vector. Table 1 shows the detailed feature extraction of different document image types. An example for Symbolic feature involves assigning color symbols or tags like ‘Red’, ‘Blue’ or ‘Green’ or ‘Magenta’ etc. for an image and storing these tags in a vector.

The feature vectors can be used to classify an image or an object. An n-dimensional vector space associated with these feature vectors is called feature space. This feature space enables visualization of feature vector and provides relationship between them. Feature Space allows us to classify an unknown sample by comparing with known samples using distance and similarity measures. Different dimensionality reduction technique such as PCA (Principal component Analysis), LDA (Linear Discriminative Analysis) can be applied to reduce the dimension of feature space.

6. Training and Testing

In a Pattern recognition and machine learning approach, the system needs to be trained before recognizing an unknown test sample. Typically, training of machine can be performed by extracting the feature values for known data samples with different classes. Recognition rate of the system can be increased by maximizing the training sets and each set consisting of more features.

The proposed system consists of seven classes with ten sample images in each class. The different afore mentioned feature parameters are extracted in order to classify the query sample.

Testing can be performed by giving an unknown sample to the trained machine for recognition. Testing involves extraction of feature parameters and comparing the extracted parameters with trained features for classification. Comparison can achieved either by performing distance or similarity measure.

7. Distance and Similarity Measure

The feature vectors of a test image or an object are used for classification. Classification is performed by comparing trained feature vectors with test feature vectors. Basically there are two methods to compare the feature vectors namely the Distance and Similarity measure. Shorter distances between two closely related vectors results in higher levels of similarity

Distance measure technique makes use of calculating the difference between the two feature vectors. If the difference is more, then the two vectors are not matching while lesser distance indicates that they match. The most widely used distance measure, metric is the Euclidean Distance technique. Given two vectors a and b then the Euclidean Distance can be given as

$$ED = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{8}$$

Similarity measure technique measures the similarity between two feature vectors by calculating the inner product between them. If the two vectors are closely matching then the similarity is more. The inner vector product between two features can be given mathematically as

$$SIM = \sum_{i=1}^{n} a_i \, b_i \tag{9}$$

The most common measure for similarity is Tanimoto metric which can be written as

$$TM = \frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} b_i^2 - \sum_{i=1}^{n} a_i b_i} \tag{10}$$

The value of TM lies between 0 and 1. If the TM equals 1, then the two feature vectors are 100% Similar.

8. Image Classification Algorithm

The statistical features such as histogram, Mean, Variance, Standard Deviation, Energy, Entropy and Skew Symmetric are calculated for known images and are tabled in a vector called as learning/training the machine. The features extracted for query document are subjected to K-NN classifier to classify the given document.

9. Priority Assignment

The test feature vectors are compared with trained feature vectors by using distance and similarity measure. The unidentified test sample is recognized as the one that belongs to the closest sample in the training set. The smallest value is considered if distance measure is used and largest value is used if similarity measure is used. This process is simple and less accurate. The accuracy can be increased by considering nearest neighbors by considering group of close feature vectors instead of selecting just a nearest training set sample. This is referred as K-Nearest Neighbor technique. K number of best matching neighbors are selected to classify the unknown sample to the given class. The value of K ranges from one to total number of images in the training set. The recognition accuracy depends on the chosen K value. As the value of K increases, we are considering matching neighbors to not matching neighbors in the training set. The test features matching with Feature space using Euclidean Distance is shown in figure 2.

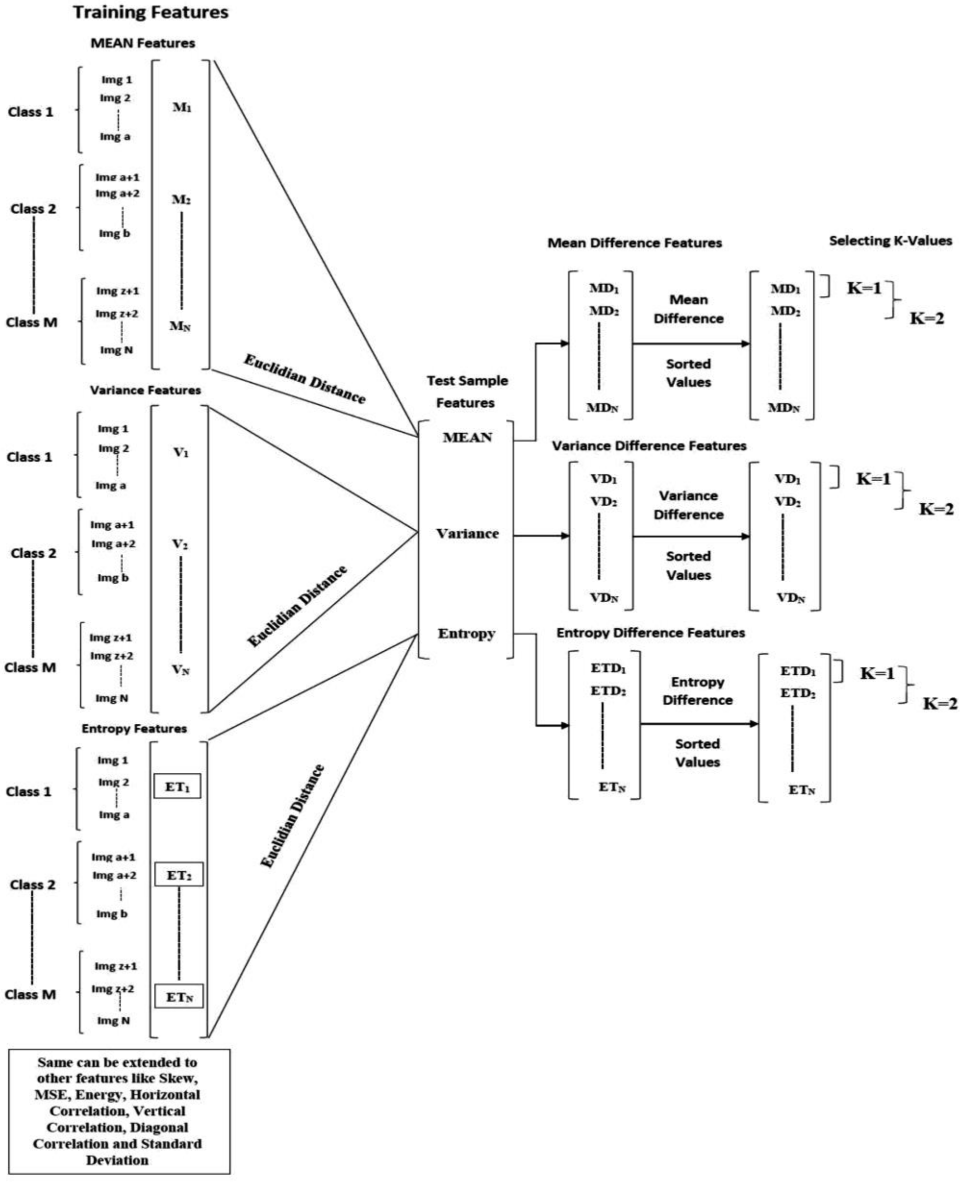


Fig. 2 Test Features matching with Feature space using Euclidean Distance.

Table 1 Image feature values for different Document Types.

Document Type	Only Text	Medical Text Label	Medical Numeric Label	Text Embedded over Picture	Text Label with More Words (News Paper)	Document with less Text (Caption)	Text Label Few words	Labelled Numeric Document	No Text (Picture)	Satellite Data
Priority	1	2	3	4	5	6	7	8	9	10
Mean	238.265	138.5	164.7	152.73	200.6	149.82	160.23	164.96	180.22	100.59
Variance	1603.20	2885.41	3497.7	3865.44	4022.01	3773.89	4679.6	4959.53	2405.9	4076.89
SD	40.0405	53.716	59.142	63.1727	63.42	61.432	68.408	70.424	49.05	63.85
Energy	0.6564	0.0095	0.0086	0.0091	0.0169	0.0112	0.0138	0.0112	0.0077	0.0069
Entropy	2.512	7.1807	7.3961	7.2349	6.835	7.1807	7.2595	7.2349	7.2531	7.5104
Skew	-2.8520	-0.675	-0.6514	-0.4492	-1.1291	-0.4895	-0.2597	-0.2880	-0.712	-0.2208
OCR with Text	1524	65	4	159	800	40	10	1	0	0
OCR with Numeric	84	1	42	15	4	12	8	38	0	0
Correlation	0.4129	0.8946	0.8573	0.9876	0.9763	0.9688	0.9756	0.9811	0.97351	0.9211

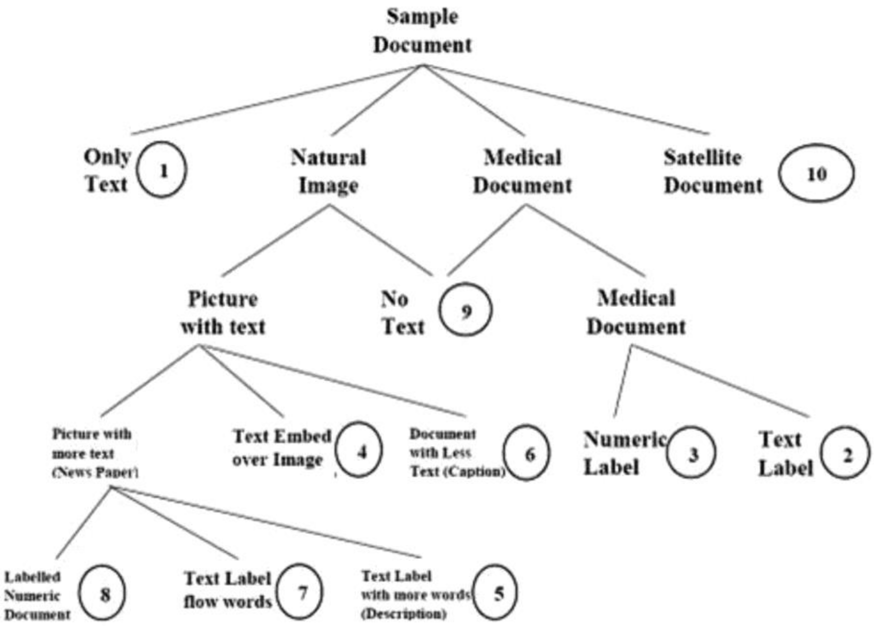


Fig. 3 Hierarchical Tree Diagram for Documents with different priority levels

A pure text document consisting of maximum text information assumed to be less random with lesser entropy and correlation, and high energy feature values are considered as the highest priority document (1). The Medical image containing textual information on the disease and patient, consisting of lesser values for Mean, Standard Deviation and Variance with a threshold range for OCR with text to be assigned the second highest priority (2). The Medical image containing numerical labelling with threshold range for OCR with numeric, is given the third highest priority (3). The textual description embedded over the picture such as certificates, the features, entropy, energy and OCR with position of the text are considered and is assigned the fourth highest priority (4). For the detailed textual information with the picture, such as the newspaper, the features, moderate entropy, energy and threshold OCR with text count are considered and treated as the fifth highest priority (5). For the picture with less text used as caption, the features entropy, energy and OCR with moderate text with its position are considered and is assigned the sixth priority (6). For a picture with few text labelling, the features entropy, OCR with minimum text and Energy are considered and treated as the seventh priority (7). For a picture with numeric labelling, the features, OCR with numeric count along with Energy is considered and is assigned eighth priority (8). For the picture image, the features, entropy, energy, Standard Deviation, Variance, correlation and OCR text/numeric count are considered and is assigned ninth priority (9). For the satellite document, the features Mean, Energy and OCR text/numeric count are considered and is the tenth priority (10). According to the above predefined priorities, the K-NN algorithm assigns the priorities for the classified Document based on its feature values are shown in figure 3.

10. Chaotic Maps

4-Dimensional Map

Hyper chaotic Lorenz map: Hyper [18] Lorenz is a 4 dimensional chaotic map represented in a differential equations having chaotic behavior for certain initial conditions. Mathematically it can be expressed as

$$\frac{dX}{dt} = p\ (Y - X) \tag{11}$$

$$\frac{dY}{dt} = c \times X + Y - X \times Z - W \tag{12}$$

$$\frac{d\,Z}{dt} = X \times Y - b \times Z \tag{13}$$

$$\frac{d\,W}{dt} = k \times X \times Y \tag{14}$$

The System exhibits chaotic behavior when the parameters are having values $p = 10$, $c = 28$ and b = 8/3.

3-Dimensional Map

Lorenz map: The Lorenz equation can be represented in differential equations having chaotic behavior for certain parameters with initial conditions. Mathematically it can be defined as

$$\frac{dX}{dt} = s\ (X - Y) \tag{15}$$

$$\frac{dY}{dt} = Y\ (r - Z) - Y \tag{16}$$

$$\frac{d\,Z}{dt} = X \times Y - b \times Z \tag{17}$$

The System exhibits chaotic behavior when the parameters are having values $s = 10$, r =28 and b = 8/3.

2-Dimensional Map

Henon map: In discrete time dynamic systems, Henon map [19] exhibit good chaotic behavior. It takes the point (X_k, Y_k) in the space and maps it to a new point.

Mathematically it can formulated as

$$X_{k+1} = Y_k - 1 + p \times X_k^2 \tag{18}$$

$$Y_{k+1} = b \times X_k \tag{19}$$

The initial valueX_0€ (0, 1) and Y_0€ (0, 1) can be used as the key for the system (X_0, Y_0). The Henon map mainly depends on two parameters. p and b, the research results shows that the value for a is 1.4 and 0.3 for b for which the Henon map exhibits chaotic nature.

Ikeda map: The 2D Ikeda map is distinct for its complicated chaotic behavior when compared to the other chaotic map. It takes the input x_i, y_i and μ in a plane and maps it to a new point. Mathematically it can be defined as

$$x_{i+1} = 1 + \mu\,(x_i \times \cos(t_n) - y_i \times \sin(t_n)) \tag{20}$$

$$y_{i+1} = \mu\,(x_i \times \sin(t_n) + y_i \times \cos(t_n)) \tag{21}$$

Where $$t_n = 0.4 - \frac{6}{(1+x_i^2+y_i^2)} \tag{22}$$

Where μ is the system parameter andx_i, y_i are the pair wise points. The system exhibits chaotic nature when μ lies in the range of [0.5 0.95]. The map depends on three values namely x_0, y_0 and μ whose corresponding initial values are $\mu = 0.9$, $x_0 = 0.1$ and $y_0 = 0.1$.

Two dimensional logistic map: The 2D logistic map is recognized well by its complicated chaotic behavior when compared to the one dimensional logistic map. It takes the input x_i, y_i and r in a plane and maps it to a new point. Mathematically it can be defined as

$$x_{i+1} = r(3y_i + 1)x_i(1 - x_i) \tag{23}$$

$$y_{i+1} = r(3x_{i+1} + 1)y_i(1 - y_i) \tag{24}$$

Where r is the system parameter x_i, y_i are the pair wise points. The map depends on three values namely x_0, y_0 and r whose corresponding initial values are $r = 1.19$, $x_0 = 0.8909$ and $y_0 = 0.3342$.

Gingerbread man map: The Gingerbread man map is a chaotic two-dimensional map. It is given by the piecewise linear transformation.

$$x_{i+1} = 1 - y_i + |x_i| \tag{25}$$

$$y_{i+1} = x_i \tag{26}$$

Where $x_0 = -0.1$ and $y_0 = 0.1$.

1-Dimensional Map

Logistic map: The basic one dimensional logistic map [20] can be formulated as

$$X_{k+1} = a \times X_k \times (1 - X_k) \tag{27}$$

Where X_k€ (0, 1). The parameter a and the intial value X_0 can be used as the key for the system (a, X_0). The results obtained from the research indicates that the system is in chaotic condition when a ranges from 3.569 <a<4.0.

Quadratic Map: The one dimensional Quadratic Chaotic map equation with initial condition r and x_0=0.1 can be mathematically defined as

$$X_{n+1} = r - X_n^2 \tag{28}$$

The System exhibits chaotic behavior when the parameters are having values $r = 1.95$.

Bernoulli map: The Bernoulli map can be mathematically given as

$$B(x_{n+1}) = \begin{cases} 2x_n & \text{If } x_n \text{ € } [0, 1/2] \\ 2x_n - 1 & \text{If } x_n \text{ € } [1/2, 1] \end{cases} \tag{29}$$

The map exhibits a chaotic behavior when x_0 = 0.2709.

Circle map: The Circle map can be mathematically can be expressed as

$$X_{n+1} = X_n + d - \left(\frac{c}{2*pi}\right) * \sin(2 \times pi \times X_n)\ mod(1) \tag{30}$$

Where$d = 0.2$, $c = 0.5$ and X_0€ [0,1]

Sine map: Sine map [21] can be defined as

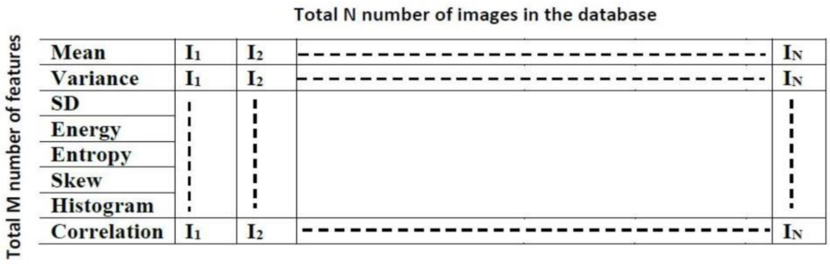
$$x_{n+1} = p \times x_n^2 \times \sin(pi \times x_n) \tag{31}$$

The system exhibits chaotic behavior when $x_0 = 0.7$ and $p = 2.3$

11. Proposed Methodology

Image Classification

In our proposed work a Document image page, containing only text, only picture, text embed over picture, labelled document, picture with more text, picture with caption, medical image documents, satellite image documents, Natural image document sets etc., are used as the input/test/query images. The document may contain a single page or more. The input sample color image is first subjected to K-NN image classification algorithm. The image features being Histogram, Mean, Variance, Standard deviation, Energy, Entropy, Skew and Correlation. The image features are arranged in a space called feature space. More number of images belonging to a single class are stored in the Database for better results. Feature Space allows us to classify an unknown sample by comparing with known samples using distance and similarity measures. The K-NN algorithm makes use of Euclidean distance measure technique to find the minimum difference between training and testing features is shown in figure 4. In order to maximize the success rate of the sample image more number of nearest neighbors are considered.



Total N number of images in the database

Total M number of features

Mean	I_1	I_2	------------------------------	I_N
Variance	I_1	I_2	------------------------------	I_N
SD	⋮	⋮		⋮
Energy	⋮	⋮		⋮
Entropy	⋮	⋮		⋮
Skew	⋮	⋮		⋮
Histogram	⋮	⋮		⋮
Correlation	I_1	I_2	------------------------------	I_N

Fig. 4 K-NN classification using sorted distance values for all different features

Mathematical model for K-NN

For a given query instance x_t, K-NN algorithm works as follows:

$$y_t = \underset{c\epsilon\{c_1,c_2,\dots\dots c_m\}}{\arg\ max} \sum_{x_i \epsilon\ N\ (x_t,k)} E(y_i,c) \tag{32}$$

Where y_t, is the predicted class for the query instance x_t , c is the class number and m is the class number present in the data. $N\ (x_t,k)$ Set of k nearest neighbors of x.
Where

$$E(a,b) = \begin{cases} 1\ if \text{ min ED} \\ 0\ if \text{ max ED} \end{cases} \tag{33}$$

Euclidean distance $ED = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$ between query instance vector a and trained vector b.

The K-NN method classification is as follows

1. Calculate the different feature values for each of the images and store them into a space called feature vector space.
2. Find the Euclidean distance measure for each feature between the sample image and the images in the database.
3. The Euclidean distance for each feature corresponding to different images are sorted in the ascending order.
4. The first column (K=1) in the figure 4 represents the best match between the sample image and the training image. The image which finds maximum occurrence in the first column is the one matching with the sample image.
5. For more accuracy consider the value of K=’n’, that is the ‘n’ nearest neighbors for classification. For the above figure 4, the classification of images belonging into the three different classes with each class containing three images is shown in figure 5.

The image so classified is subjected to check if it contains the text. The text primarily carries the confidential information in the image. The magnitude of text contained in the image represents the density of confidential information. Hence the presence of text with its magnitude is to be identified. The OCR method checks if the image contains the text. The magnitude of the text maybe a few words or more (1000s of words). The OCR also provides the location of the text in the image. Based on the density of text and its location, the documents are classified further and each image is assigned with a priority value as shown in the figure 3.

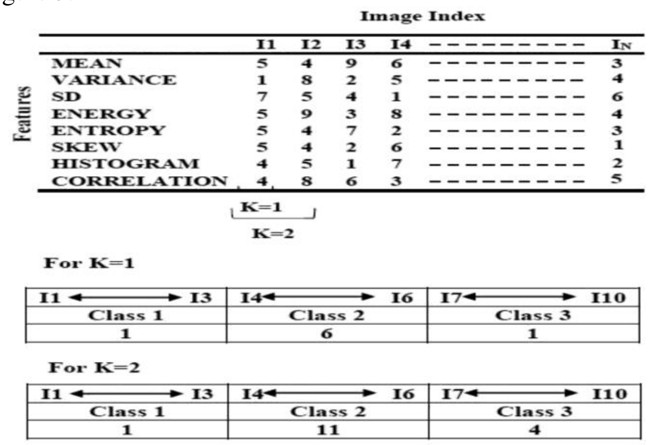


Image Index

Features	I1	I2	I3	I4	----------	IN
MEAN	5	4	9	6	----------	3
VARIANCE	1	8	2	5	----------	4
SD	7	5	4	1	----------	6
ENERGY	5	9	3	8	----------	4
ENTROPY	5	4	7	2	----------	3
SKEW	5	4	2	6	----------	1
HISTOGRAM	4	5	1	7	----------	2
CORRELATION	4	8	6	3	----------	5

For K=1

I1 ⟷ I3	I4 ⟷ I6	I7 ⟷ I10
Class 1	Class 2	Class 3
1	6	1

For K=2

I1 ⟷ I3	I4 ⟷ I6	I7 ⟷ I10
Class 1	Class 2	Class 3
1	11	4

Fig. 5 Classification of images belonging into the three different classes with each class containing three or four images

Encryption

Encryption is a process of converting an intelligent form of information to an unintelligent form. The confusion and diffusion procedures are followed for the encryption. The confusion and diffusion techniques are used with the multidimensional chaotic maps. The priority levels of the documents determines the levels of security for each document. To increase the level of security, the block size of the image, the chaotic maps used for confusion and diffusion, and the technique used for establishing the interdependency between neighboring pixels plays a very important role. Based on the required priority levels of documents, the block size, the maps and method of interdependency are chosen.

Confusion and Diffusion

The confusion is a process of scrambling the Document image pixels/blocks. The diffusion is a process of modifying the values of pixels and establishing interdependency among the neighboring pixels. The detailed confusion and diffusion process are described in Table 2.

Decryption

Decryption is a process in which a plain image is extracted from the given cipher image and is a reverse process of encryption. For the given cipher image, the priority level is extracted from the key. Based on this priority level, the inverse second level diffusion is performed first and then the inverse first level diffusion is performed using the corresponding map from the key. This resultant image is subjected to the inverse process of confusion using the corresponding chaotic map from the key.

Table 2 Encryption of different documents with different priorities.

Document Image Page Type	Priority P_r	Confusion map	Block Size	First level Diffusion map	Key	Encryption Technique
Only Text	1	4D Hyper Lorenz map X=0.0000000000778899 Y=0.0000000000874533 Z=0.0000000000898447 W=0.000000000098876 p=10 c=28 b=8/3	1x1 pixels	1D Logistic map a= 3.9 X_0= 0.1	K={P_r ,X, Y, Z, W, p, c, b, a, X_0} P_r=1 X=0.00000000007788 99 Y=0.00000000008745 33 Z=0.00000000008984 47 W=0.0000000000988 76 p=10 c=28 b=8/3 a= 3.9 X_0= 0.1	Confusion 1. Divide the image of size 512×512 to equal number of blocks of size 1×1 pixels. 2. Generate the chaotic sequence of length $\frac{512}{1}\times\frac{512}{1}$ using the 4D Hyper Lorenz map, convert them into integers and obtain the remainder using modulus as the unique index values. 3. Permute the blocks according to the sequence generated by the 4D Hyper Lorenz map in step 2. Diffusion *First level diffusion* 1. Generate the chaotic sequence of size 512 × 512 using the 1D logistic map. 2. The generated sequence is converted into integer by multiplying with a factor of 10^{15} and obtain the remainder by using Mod 255. 3. Modify the confused pixels by XORing them with the chaotic sequence generated in step 2. *Second level diffusion* 1. The 3x3 Kernel is traversed on the entire image and the Mean is calculated. 2. The obtained Mean value is XORed with every element within the Kernel. [15].
Text labelled Medical image	2	3D Lorenz map X=0.0000000000856672 Y=0.0000000000785563 Z=0.0000000000889732 s=10 r=28 b=8/3	2x2 Pixels	1D Logistic map a= 3.9 X_0= 0.1	K= { $P_{r,}$ X, Y, Z, s,r,b,a,X_0} $P_{r=}$2 X=0.00000000008566 72 Y=0.00000000007855 63 Z=0.00000000008897 32 s=10 r=28 b=8/3 a= 3.9 X_0= 0.1	Confusion 1. Divide the image of size 512×512 to equal number of blocks of size 2×2 pixels. 2. Generate the chaotic sequence of length $\frac{512}{2}\times\frac{512}{2}$ using the 3D Lorenz map, convert them into integers and obtain the remainder using modulus as the unique index values. 3. Permute the blocks according to the sequence generated by the 3D Lorenz map in step 2. Diffusion *First level diffusion* 1. Generate the chaotic sequence of size 512 × 512 using the logistic map. 2. The generated sequence is converted into integer by multiplying with a factor of 10^{15} and obtain the remainder by using Modulus 255. 3. Modify the confused pixels by XORing them with the chaotic sequence generated in step 1. *Second level diffusion* 1. Generate the Fibonacci series of length equal to total number of image pixels. 2. Generated Series are XORed with image pixels in both forward and reverse directions.

Numeric labelled Medical image	3	2D Henon map X=0.6315477 Y= 0.18906343 p=1.4 b=0.3	4x4 pixels	1D Logistic map a= 3.9 X_0= 0.1	K= {P_r ,X, P,b,a,X_0} $P_{r=}$3 X=0.6315477 Y= 0.18906343 p=1.4 b=0.3 a= 3.9 X_0= 0.1	Confusion 1. Divide the image of size 512×512 to equal number of blocks of size 4×4 pixels. 2. Generate the chaotic sequence of length $\frac{512}{4}\times\frac{512}{4}$ using the 2D Henon map, convert them into integers and obtain the remainder using modulus as the unique index values. 3. Permute the blocks according to the sequence generated by the 2D Henon map in step 2. Diffusion *First level diffusion* 1. Generate the chaotic sequence of size 512 × 512 using the logistic map. 2. The generated sequence is converted into integer by multiplying with a factor of 10^{15} and obtain the remainder by using Modulus 255. 3. Modify the confused pixels by XORing them with the chaotic sequence generated in step 1 *Second level diffusion* 1. Generate the two random sequence using 1D Logistic map of length equal to total number of rows and total number of columns in the image. 2. XOR the obtained sequence with pixels in both row as well as column. 3. Establish interdependency within an image by XORing its previous row/column with current row/column.
Picture with more Text	4	2D Ikeda map X=0.1 Y= 0.1 μ = 0.9	8x8 pixels	1D Logistic map a= 3.9 X_0= 0.1	K={P_r,X,Y, μ,a,X_0} P_r=4 X=0.1 Y= 0.1 μ = 0.9 a= 3.9 X_0= 0.1	Confusion 1. Divide the image of size 512×512 to equal number of blocks of size 8×8 pixels. 4. Generate the chaotic sequence of length $\frac{512}{8}\times\frac{512}{8}$ using the 2D Ikeda map, convert them into integers and obtain the remainder using modulus as the unique index values. 2. Permute the blocks according to the sequence generated by the 2D Ikeda map in step 2. Diffusion *First level diffusion* 1. Generate the chaotic sequence of size 512 × 512 using the logistic map. 2. The generated sequence is converted into integer by multiplying with a factor of 10^{15} and obtain the remainder by using Modulus 255. 3. Modify the confused pixels by XORing them with the chaotic sequence generated in step 1. *Second level diffusion* 1. Scan the pixels in the image in the square wave path for both row as well as column. 2. Perform XOR between the pixels which comes on that path.

Text label with more Description	5	2D Logistic map X=0.8909 Y= 0.3342 r=1.19	16x16 pixels	1D Logistic map a= 3.9 X_0= 0.1	K={ P_r,,X,Y,r,a, X_0} P_r=5 X=0.8909 Y= 0.3342 r=1.19 a= 3.9 X_0= 0.1	Confusion 1. Divide the image of size 512×512 to equal number of blocks of size 16×16 pixels. 2. Generate the chaotic sequence of length $\frac{512}{16}\times\frac{512}{16}$ using the 2D Logistic map, convert them into integers and obtain the remainder using modulus as the unique index values. 3. Permute the blocks according to the sequence generated by the 2D Logistic map in step 2. Diffusion *First level diffusion* 1. Generate the chaotic sequence of size 512 × 512 using the logistic map. 2. The generated sequence is converted into integer by multiplying with a factor of 10^{15} and obtain the remainder by using Modulus 255. 3. Modify the confused pixels by XORing them with the chaotic sequence generated in step 1. *Second level diffusion* 1. Scan the pixels with in the image in the triangular wave path for both row as well as column. 2. Perform XOR between the pixels which comes on that path.
Text with Caption	6	2D Gingerbread man map X_0=0.1 Y_0=0.1	32x32 pixels	1D Logistic map a= 3.9 X_0= 0.1	K= {P_r, X_{0r}, Y_0, a,X_0} P_r=6 X_0=0.1 Y_0=0.1 a= 3.9 X_0= 0.1	Confusion 1. Divide the image of size 512×512 to equal number of blocks of size 32×32 pixels. 2. Generate the chaotic sequence of length $\frac{512}{32}\times\frac{512}{32}$ using the 2D Gingerbread man map, convert them into integers and obtain the remainder using modulus as the unique index values. 3. Permute the blocks according to the sequence generated by the 2D Gingerbread man map in step 2. Diffusion *First level diffusion* 1. Generate the chaotic sequence of size 512 × 512 using the logistic map. 2. The generated sequence is converted into integer by multiplying with a factor of 10^{15} and obtain the remainder by using Modulus 255. 3. Modify the confused pixels by XORing them with the chaotic sequence generated in step 1. *Second level diffusion* 1. Scan the pixels within the image in the raster scan path order. 2. Perform XOR between the pixels which comes on that path.

Text label with few Description	7	1D Quadratic map x_0=0.1 r = 1.95	64x64 pixels	1D Logistic map a= 3.9 X_0= 0.1	K={ P_r, x_0,r,a,X_0} P_r=7 x_0=0.1 r = 1.95 a= 3.9 X_0= 0.1	Confusion 1. Divide the image of size 512 × 512 to equal number of blocks of size 64 × 64. 2. Generate the chaotic sequence of length $\frac{512}{64}\times\frac{512}{64}$ using the 1D Quadratic map, convert them into integers and obtain the remainder using modulus as the unique index values. 3. Permute the blocks according to the sequence generated by the 1D Quadratic map in step 2. Diffusion *First level diffusion* 1. Generate the chaotic sequence of size 512 × 512 using the logistic map. 2. The generated sequence is converted into integer by multiplying with a factor of 10^{15} and obtain the remainder by using Modulus 255. 3. Modify the confused pixels by XORing them with the chaotic sequence generated in step 1. *Second level diffusion* 1. Divide the image into triangular shaped four equal units. 2. Perform XOR operation between every individual unit with remaining three units.
Text label with Numeric	8	1D Bernoulli's map x_0=0.2709	128x128 pixels	1D Logistic map a= 3.9 X_0= 0.1	K= { P_r,x_0,a,X_0} P_r=8 x_0=0.2709 a= 3.9 X_0= 0.1	Confusion 1. Divide the image of size 512×512 to equal number of blocks of size 128×128 pixels. 2. Generate the chaotic sequence of length $\frac{512}{128}\times\frac{512}{128}$ using the 1D Bernoulli's map, convert them into integers and obtain the remainder using modulus as the unique index values. 3. Permute the blocks according to the sequence generated by the 1D Bernoulli's map in step 2. Diffusion *First level diffusion* 1. Generate the chaotic sequence of size 512×512 pixels using the logistic map. 2. The generated sequence is converted into integer by multiplying with a factor of 10^{15} and obtain the remainder by using Modulus 255. 3. Modify the confused pixels by XORing them with the chaotic sequence generated in step 1. *Second level diffusion* 1. Divide the image into triangular shaped four equal units. 2. Perform XOR operation between every individual unit with remaining three units. 3. Scan the pixels with in the image in the zig-zag path order. 4. Perform XOR between the pixels which comes on that path.

Image with no Text	9	Circle map X=0.1 d = 0.2 c=0.5	256x256 pixels	1D Logistic map a= 3.9 X_0= 0.1	K={ P_r ,X,d,c,a,X_0} P_r=9 X=0.1 d = 0.2 c=0.5 a= 3.9 X_0= 0.1	Confusion 1. Divide the image of size 512 × 512 to equal number of blocks of size 256 × 256. 4. Generate the chaotic sequence of length $\frac{512}{256}\times\frac{512}{256}$ using the 1D Circle map, convert them into integers and obtain the remainder using modulus as the unique index values. 2. Permute the blocks according to the sequence generated by the 1D Circle map in step 2. Diffusion *First level diffusion* 1. Generate the chaotic sequence of size 512×512 pixels using the logistic map. 2. The generated sequence is converted into integer by multiplying with a factor of 10^{15} and obtain the remainder by using Modulus 255. 3. Modify the confused pixels by XORing them with the chaotic sequence generated in step 1. *Second level diffusion* 1. Scan the pixels with in the image in the bottom to top approach order. 2. Perform XOR between the pixels which comes on that path.
Satellite image Document	10	Sine map x_0=0.7 p =2.3	512x512 pixels	1D Logistic map a= 3.9 X_0= 0.1	K={ P_r ,x_0,p, a,X_0} P_r=10 x_0=0.7 p =2.3 a= 3.9 X_0= 0.1	Confusion 1. Divide the image of size 512×512 to equal number of blocks of size 512×512 pixels. 2. Generate the chaotic sequence of length $\frac{512}{512}\times\frac{512}{512}$ using the 1D Sine map, convert them into integers and obtain the remainder using modulus as the unique index values. 3. Permute the blocks according to the sequence generated by the 1D Sine map in step 2. Diffusion *First level diffusion* 1. Generate the chaotic sequence of size 512×512 pixels using the logistic map. 2. The generated sequence is converted into integer by multiplying with a factor of 10^{15} and obtain the remainder by using Modulus 255. 3. Modify the confused pixels by XORing them with the chaotic sequence generated in step 1. *Second level diffusion* 1. Scan the pixels with in the image in the right to left direction path order. 2. Perform XOR between the pixels which comes on that path.

12. Assessment Parameters

To assess the quality of the encryption method followed, the statistical and dynamic assessment parameters are calculated and compared with their ideal values. The statistical parameters include MSE, PSNR, correlation, entropy, key sensitivity, key space, SSIM and UIQ, whereas the dynamical assessment parameters include NPCR and UACI. The encryption time taken for the set of Documents when encrypted without classification by using single common encryption algorithm and Documents classified according to their type and encrypted using different algorithms with multi-dimensional chaotic maps and diffusion technique is determined. The complexity of the algorithm is varied by using the different multi- dimensional chaotic maps.

Histogram

Histogram is a graphical representation of the distribution of number of pixels for a particular intensity level. The histogram for an encrypted image should be flat or uniform distribution for all the pixel intensity levels. A flat histogram depicts the difficulty in understanding/predicting the plain image. It is desirable to have uniform histograms for two cipher images which are obtained from the same plain image but with a tiny change in the key value. The variances of the histograms are determined and is tabulated in Table 3 for text document image. The smaller value of variance depicts higher uniformity. The variance of the histograms are calculated by using the equation

$$var(h) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(h_i - h_j)^2}{n^2} \tag{34}$$

Where h_i and h_j the number of pixels which grey values equals to i and j.

Table 3 Histogram Variance obtained for 5% change in initial parameters

Key with parameters	X_0	Y_0	Z_0	W_0	x_0
Text Document Image	0.7365	0.7811	0.7565	0.7617	0.7250

Tests have been conducted to check if the ciphered image produced with different keys is producing the flat histogram with a uniform variance. For example, the variance values of the histogram obtained for the plain image with highest priority using the key K1 (X_0 =0.0000000000778899, Y_0 =0.0000000000123654, Z_0 =0.00000000000657789, $W_0 = 10^{-15}, a = 10, c = 28$ and b = 8/3 and $x_0 = 0.1$,$r = 3.97$), and the histogram obtained for a ciphered image with different keys which are obtained for 5% change in each of the initial parameters of K1 are calculated. The results obtained are tabulated in Table 3. The uniformity in the variance for different keys extracted out of uniform change in different parameters reveal that the ciphering technique is key-sensitive and resistant to statistical attacks.

Ciphering Time

The proposed ciphering algorithm is implemented using MATLAB 2014 Software on Intel core i3 processor having 2GB RAM and 500GB HD. The encryption time taken for all types of document images with and without priority assignments are tabulated in Table 4. Document containing different information are encrypted with different level of complexity. Here the prioritized documents are encrypted with different block sizes as indicated in Table 2, based on their image features, the total encryption time taken for a bunch of documents is less. On the other hand, when a bunch of documents are encrypted without assigning priorities where all the documents are using a common algorithm with same chaotic map, same block size and the same method of diffusion, leading to more encryption time. It is observed that encryption of a set of Documents experimented without assigning any priority has taken 4630.774422 seconds and that with priority, has taken 658.443416 seconds. The Encryption time versus priority is graphically shown in figure 6. To further reduce the encryption time, the high dimensional chaotic maps which require lesser time to generate the required set of sequences are used. The high dimensional chaotic sequences are more aperiodic when compared to lower dimensional chaotic maps and thereby increases the security of the system. The encryption time taken for a Lena image of size 512 x 512 in the proposed method is compared with the time taken for other methods in Table 5. It is observed in Table 5 that the proposed method has taken significantly lesser time for encryption.

Table 4 Execution Time obtained for different Document images with and without Priority

Image	Encryption Time (seconds)
Proposed (Lena 512 x 512)	0.020298
[24]	0.130
[25]	0.175
[26]	0.125

Table 5 Encryption time Comparison for different existing methods

	Time(sec) with priority	Time(sec) without priority
First	465.2658	465.2658
Second	170.970537	452.502153
Third	19.759614	444.861688
Fourth	2.228033	449.647004
Fifth	0.095617	447.941053
Sixth	0.035254	492.808793
Seventh	0.027551	461.581642
Eighth	0.023044	447.308014
Ninth	0.020298	512.699405
Tenth	0.017668	456.15887
Total	658.443416	4630.774422

Document containing different information are encrypted with different level of complexity. Here the prioritized documents are encrypted with different block sizes as indicated in Table 2, based on their image features, the total encryption time taken for a bunch of documents is less. On the other hand, when a bunch of documents are encrypted without assigning priorities where all the documents are using a common algorithm with same chaotic map, single block size and the same method of diffusion, leading to more encryption time. It is observed that encryption of a set of Documents experimented without assigning any priority has taken 4630.774422 seconds and that with priority has taken 658.443416 seconds. The Encryption time versus priority is graphically shown in figure 6. To further reduce the encryption time, the high dimensional chaotic maps which require lesser time to generate the required set of sequences are used. The high dimensional chaotic sequences are more aperiodic when compared to lower dimensional chaotic maps and thereby increases the security of the system. The encryption time taken for a Lena image of size 512 x 512 in the proposed method is compared with the time taken for other methods in Table 5. It is observed in Table 5 that the proposed method has taken significantly lesser time for encryption.

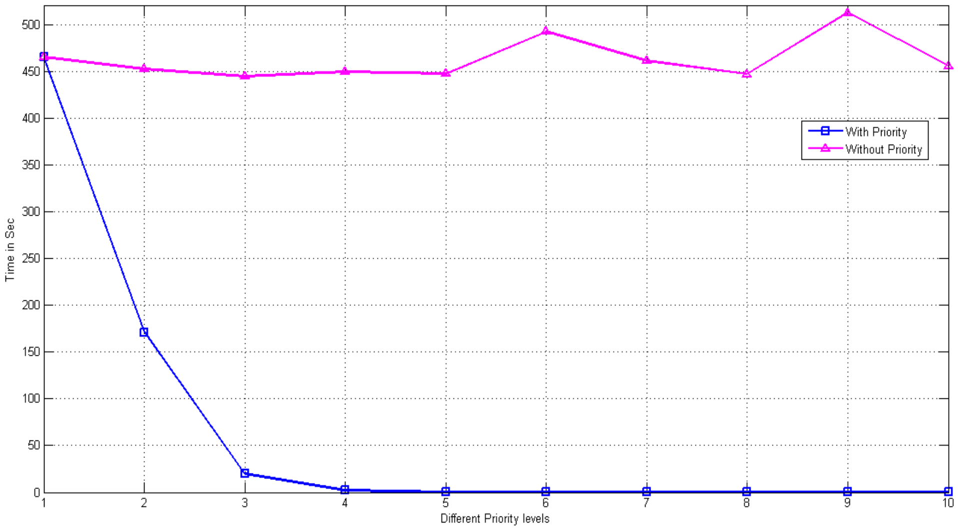


Fig. 6 Graphical comparison of utilizing system resource for encryption by Document images with and without priority.

Mean Squared Error and Peak Signal to Noise Ratio

The Peak Signal to Noise Ratio (PSNR) is used to assess the encryption scheme. It represents the amount of noise content present in the ciphered image. The PSNR is calculated using the equation

$$PSNR = 10log_{10}(\frac{255^2}{MSE}) \tag{35}$$

$$\text{Where } MSE = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}[plain(i,j)-cipher(i,j)]^2}{M*N} \tag{36}$$

The value of PSNR>30 dB is not advisable as it is possible to extract the original information from the ciphered image. The desirable value is less than 8dB. The result yields 4.772dB as depicted in Table 6, this indicates the difficulty in reconstructing the plain image from the cipher image. PSNR also represents the distortion of the plain image when subjected to encryption. It is observed that the MSE is more, for a small change in the initial conditions. More the MSE, the better is the algorithm. A small value of MSE enables the interceptor to visualize the original image. These parameters contribute to confidentiality of the document.

Entropy Analysis

The information entropy measures the randomness in the image. It is calculated by the equation

$$\text{ET (m)} = - \sum_{i=0}^{L-1} p(m_i) \text{ x } log_2(p(m_i)) \qquad (37)$$

For a gray scale image with 2^8 states of information, if all the 256 states appear with the same probability the entropy value is equal to 8. The experimental result yields an entropy very close to 8 as shown in Table 6. This is possible only when all pixels in the cipher image appear with same probability (equally probable). This indicates that the cipher image is random in pixel distribution. When all the pixels in the ciphered image appear with equal probability, it is very difficult to predict the original image by taking the statistical analysis. This parameter contribute to unpredictability and degree of uncertainty of the document.

Correlation Analysis

Correlation is a measure of relationship between the plain and cipher images. It checks if they are similar or not. When a plain image is encrypted, the pixel positions are interchanged and their values get modified. Hence the pattern of the ciphered image is different from that of the plain image. This can be measured by calculating the covariance between the set of pixels in different directions both in the plain and cipher images. For a set of pixels, the correlation coefficient of '1' indicates that the images are similar and a '0' indicates the dissimilarity between them. The equations for calculating the correlation is given by

$$C_{xy} = \frac{COVR(x,y)}{\sqrt{V(x)}\sqrt{V(y)}} \qquad (38)$$

Where $COVR(x, y)$ is the Covariance between x and y can be formulated as

$$COVR(x,y) = \frac{1}{n}\sum_{i=1}^{n} E((x_i - \mu(x))(y_i - \mu(y))) \qquad (39)$$

Where, x and y are two adjacent pixels values in the image, V (x) is the variance of variable x,

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu(x))^2 \qquad (40)$$

$$\mu(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (41)$$

For a sample of about 2000 pixels are taken in horizontal, vertical and diagonal directions and the results are listed in Table 6. The results indicate that there is very high dissimilarity between the plain and ciphered images. The results indicate that there is a large randomness in the cipher image and very high dissimilarity between the plain and ciphered images. This contributes to reduction in the redundancy as low as possible in the cipher image.

Table 6 Encryption results obtained for different document images with different priority

Priority	PSNR	MSE	HORI	VERTI	DIAG	ENTROPY	NPCR	UACI	SSIM	UIQ
First	4.772	$2.17X10^4$	-0.02619	-0.0308	-0.00996	7.9993	100	33.465	$-7.75X10^{-4}$	$-7.74X10^{-4}$
Second	7.958	$9.127X10^3$	-0.0119	-0.0175	-0.0044	7.9992	99.609	33556	$7.00X10^{-4}$	$6.99X10^{-4}$
Third	7.044	$1.28X10^4$	-0.0057	0.01109	-0.0214	7.9992	99.616	33.285	$7.701X10^{-4}$	$7.692X10^{-4}$
Fourth	6.409	$1.48X10^4$	-0.0055	-0.0171	0.00442	7.9993	99.59	33.49	0.001033	0.001032
Fifth	6.593	$8.417X10^3$	-0.022	0.00597	-0.0157	7.9993	99.57	32.92	$-1.6699X10^{-4}$	$-1.682X10^{-4}$
Sixth	7.931	$1.014X10^4$	0.0194	-0.004	0.0089	7.9992	99.56	32.89	$-1.023X10^{-4}$	$-1.034X10^{-4}$
Seventh	8.368	$9.4663X10^3$	-0.005	-0.0153	0.012	7.9993	99.43	31.93	$-3.911X10^{-4}$	$-3.921X10^{-4}$
Eighth	9.515	$7.27X10^3$	-0.0154	0.0043	-0.0024	7.9992	99.334	30.45	$7.107X10^{-4}$	$7.093X10^{-4}$
Ninth	9.342	$7.50X10^3$	0.016	0.0121	0.0209	7.9993	99.24	30.19	-0.001339	-0.001335
Tenth	9.367	$6.83X10^3$	0.0117	0.02184	0.0266	7.9992	99.21	30.08	-0.001633	-0.001635

Key sensitivity and key space analysis

Key sensitivity is a test to check how many pixels are changing in their values for a tiny change in the original encryption key. For a good encryption scheme, two keys which differ in one bit produce significantly different ciphers. The Key value depends on the initial parameters and the control parameters of the maps being used for encryption. Also a bit change in the key at the receiving end will produce completely different plain images. The key space represents the total number of different key values for the given precision. The key space should be large enough to make it difficult for the intruder to crack the correct key to reconstruct the plain image. It should take years of time for the successful brute force attack. The keys used for the encryption of each class of document image is different from the other class. The key space is depending on the key value which basically depends on the initial parameters used, the precision of the real pseudorandom numbers being generated and the number of iterations the algorithm is repeated for ciphering. For example, in the encryption of highest priority image, the key used is 4D Hyper Chaotic Lorenz map with initial conditions X_0,Y_0,Z_0, W_0 and Step size h and the 1D logistic map with initial condition of x_0 and the control parameter r is being used. The precision of the random number sequences being generated is equal to 10^{-15} [22]. The length of the Key is equal to 10^{-16} times the number of initial conditions. That is $(10)^{(16)^7}$ which is greater than 2^{318}. The size of the key space should be above 2^{100} [23] to get rid of brute force attacks. It is cleared from the results obtained for the proposed scheme that the algorithm is resistant to brute force attacks.

NPCR and UACI

NPCR and UACI are the assessment parameters in respect of differential attack for the cipher image. NPCR denotes the rate at which the number of pixels changes in their values. A change is reflected as 1 and no change reflected as 0 for a single bit change in the cipher image. The UACI denotes the unified average change intensity, that is, the average change in the intensity values of pixels of the ciphered images obtained when a bit change is made at the plain image. Ideally the NPCR should be 100% and the UACI should be 33.465%. The NPCR and UACI represents the strength of the encryption. It is cleared from the results obtained for the proposed scheme that the algorithm is resistant to brute force attacks. The results are tabulated in the Table 6. The equations for calculating the NPCR and UACI are given below.

$$\text{NPCR} = \frac{\sum_{i,j} W(i,j)}{(MXN)} \times 100 \tag{42}$$

Where M and N are the width and height of the image.
$W(i,j)$ Can be defined as

$$W(i,j) = \begin{cases} 1 & Cip1(i,j) \neq Cip2(i,j) \\ 0 & Cip1(i,j) = Cip2(i,j) \end{cases} \tag{43}$$

$Cip1(i,j)$ Grey value of cipher image and $Cip2(i,j)$ Grey value of new cipher image

Universal Image Quality Index (UIQ)

Let x= $\{x_i|i = 1,2, \dots \dots .. M \times N\}$ and y= $\{y_i|i = 1,2, \dots \dots .. M \times N\}$ be the original and cipher images respectively, then the quality index can be defined as

$$UIQ = \frac{4\times\sigma_{x,y}\times x\times y}{(\sigma_x^2+\sigma_y^2)\times[(\bar{x})^2+(\bar{y})^2]} \tag{44}$$

Where

$$\bar{x} = \frac{1}{M\times N}\sum_{i=1}^{M\times N} x_i \tag{45}$$

$$\bar{y} = \frac{1}{M\times N}\sum_{i=1}^{M\times N} y_i \tag{46}$$

$$\sigma_x^2 = \frac{1}{M\times N}\sum_{i=1}^{M\times N}(x_i - \bar{x})^2 \tag{47}$$

$$\sigma_y^2 = \frac{1}{M\times N}\sum_{i=1}^{M\times N}(y_i - \bar{y})^2 \tag{48}$$

$$\sigma_{x,y} = \frac{1}{M\times N}\sum_{i=1}^{M\times N}(x_i - \bar{x}) \times (y_i - \bar{y}) \tag{49}$$

The ideal value for UIQ is -1. The results obtained for different priority Documents are tabulated in Table 6.

Structural Similarity Index Measure (SSIM)

The SSIM is a quality metric for images. It measures the similarities between the two images with a reference image. The reference image being the uncompressed or noise free image. This metric compares the contrast, luminance and structural information between equal sized gray level images. Its value is in the range of -1 and 1. A, 1 means that the two images are exactly the same but a -1 means they are dissimilar. Mathematically it can be given as

$$SSIM = \frac{(2\times\bar{x}\times\bar{y}+c1)\times(2\sigma_{x,y}+c2)}{(\bar{x}^2+\bar{y}^2+c1)\times(\sigma_x^2+\sigma_y^2+c2)} \tag{50}$$

Where $c1$ and $c2$ are constants. The values of SSIM obtained for the different priority images are tabulated in Table 6.

Comparison of security levels

The Training Database contains 10 different classes of Document images obtained from internet source. Each class contains a set of 7 images exhibiting all the attributes resulting in totally 70 images for training the system as shown in Figure 7. For Each image the different features are extracted and stored them in MATLAB library called '.mat' file. This file is loaded back when querying the test image for recognizing the Document type. Encryption time Comparison for different existing methods are detailed in Table 5. The results obtained for a set of documents of size 512x512 in the proposed method of encryption are listed in Table 6. It is observed that the number of pixel change rate and unified average change intensity of images are equal to the ideal values in the proposed method. It indicates that the proposed method of encryption is 100% resistant to dynamical attacks. The entropy of the proposed method is greater than that of the existing method and indicates that the proposed method yields more randomness and hence the leakage of information is negligible. The results of the proposed method for other document types with different block sizes are observed in Table 6. The encryption of different document image types for 4×4 pixels are compared in Table 7. The Comparison of Proposed 4×4 pixels Encryption results with Traditional block cipher Techniques are depicted in Table 8. The Different Document images when encrypted with priority, the corresponding cipher images obtained with their histogram and correlation in three different directions namely Horizontal, Vertical and Diagonal are shown in Figure 8.

Table 7 Comparison of Encryption results obtained for different document images

TYPE	HORI	VERTI	DIAG	NPCR	UACI	ENTROPY	PSNR	MSE	UIQ	SSIM
Proposed (4 x 4 pixels)	-0.0262	-0.031	-0.009	100	33.465	7.9993	4.772	$2.17\mathrm{x}10^4$	$-7.74\mathrm{x}10^{-4}$	$-7.75\mathrm{x}10^{-4}$
[24]	0.0603	-0.0692	0.0487	99.66	33.49	NA	NA	NA	NA	NA
[25]	0.0079	0.0038	0.007	99.545	33.42	7.997	NA	NA	NA	NA
[26]	-0.0702	-0.0782	0.0039	NA	NA	NA	NA	NA	NA	NA
[27]	0.0004	0.0006	NA	99.61	33.47	7.997	NA	NA	NA	NA

Table 8 Comparison of Proposed Encryption results with Traditional block cipher Techniques

TYPE	HORI	VERTI	DIAG	NPCR	UACI	ENTROPY	PSNR	MSE	UIQ	SSIM
Proposed (4 x 4 pixels)	-0.0057	0.01109	-0.0214	99.616	33.285	7.9992	7.044	$1.28\mathrm{x}10^4$	$-7.692\mathrm{x}10^{-4}$	$-7.701\mathrm{x}10^{-4}$
[28]	0.0208	0.0021	0.00012	NA	NA	7.9999	NA	NA	NA	NA
[29]	-0.0269	0.01096	0.0510	99.605	33.441	7.999437	NA	NA	NA	NA
[30]	0.030768	0.019044	0.010293	99.6058	33.4648	7.99769	NA	NA	NA	NA

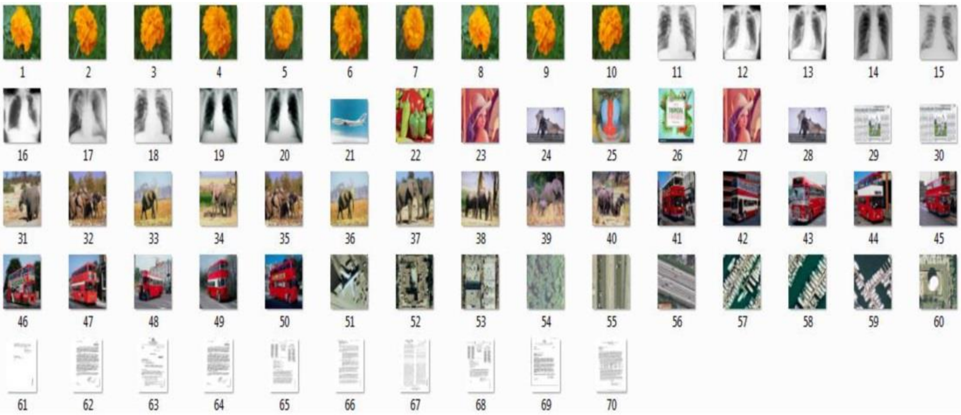


Fig. 7 Total number of different Document images used for training database.

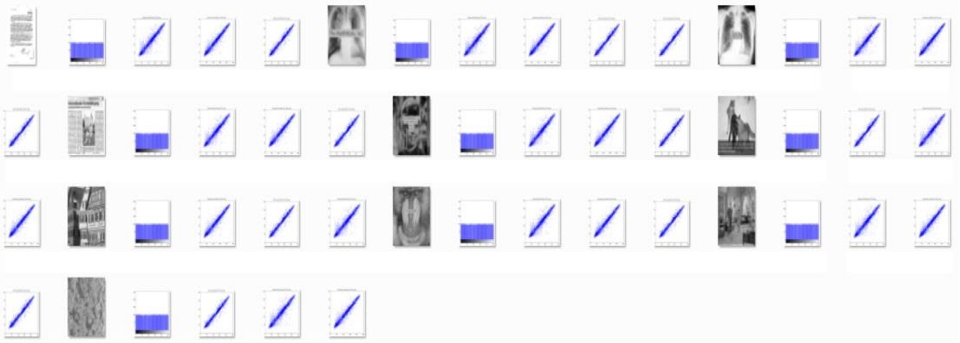


Fig. 8 Different Document images encrypted with different priority and Histogram of cipher images with its correlation in three different directions namely Horizontal, Vertical and Diagonal.

Table 9 Test result by NIST for the sequence generated by Different Dimensional Chaotic maps

	P-Value				
Statistical Analysis	4D chaotic Map	3D chaotic Map	2D chaotic Map	1D chaotic Map	Status
Mono Bit Frequency Test	0.929568022278	0.573775403633	0.426325543384	0.511663282696	Success
Block Frequency Test	0.914496908439	0.774138391129	0.220220646602	0.364406710571	Success
Run Test	0.035318327624	0.038169651362	0.036460364749	0.098222077056	Success
Longest Run Ones	0.749822008129	0.968200174367	0.200825122695	0.877325270843	Success
Binary Matrix Rank Test	0.498218033841	0.291891444943	0.085200615631	0.058955719226	Success
Spectral Test	0.935353907956	0.491297124216	0.655518578368	0.528110313792	Success
No over Lapping Template Matching Test	0.999252260332	0.623258576742	0.412836786198	0.929710883665	Success
Overlapping Template Matching Test	0.853100388979	0.270571775094	0.488415602427	0.786016922447	Success
Universal Statistic Test	0.061368829139	0.319352527457	0.294836833019	0.579319917432	Success
Linear Complexity Test	0.808840178441	0.423178207016	0.919679104285	0.959466390424	Success
Serial Test	0.999998513560	0.999999999999	1.000000000000	1.000000000000	Success
Approx. Entropy Test	1.000000000000	0.999869306864	0.977601055158	0.543961589891	Success
Cumulative Sums Test Forward	0.984155397448	0.961531188055	0.536609751712	0.831316404987	Success
Cumulative Sums Test Reverse	0.997700313205	0.629222570292	0.547547656132	0.301119661875	Success
Random Excursion Test	0.471203771279	0.737055710199	0.762173877605	0.921218168648	Success
Random Excursions Variant Test	0.842342484558	0.980593202196	0.625123538855	0.423310463480	Success

NIST Test Analysis

There are different complexity measurement techniques to measure the randomness of a given chaotic sequence. In this paper a NIST (National Institute of Standards and Technology) Analysis has been conducted to quantitatively estimate the complexity of different dimensional (4D, 3D, 2D and 1D) chaotic maps. The complexity of the proposed scheme can be assessed by making use of NIST special publication 800-22 (SP 800-22) [17]. There are 16 different statistical test of special publication SP 800-22 [16]. The different statistical methods are 1. Mono bit test, 2. Frequency test within block 3.Runs test, 4. Longest run ones test, 5. Binary matrix rank test, 6. Spectral test, 7. Non overlapping template matching test, 8. Overlapping template matching test, 9. Universal statistical test, 10. Lempel-Ziv compression test, 11. Linear Complexity test, 12. Serial test, 13. Approximate Entropy test, 14. Cumulative sums test, 15. Random excursion test and 16. Random excursion variant test. For each of these tests the value of P is calculated from a binary sequence generated by the multi-dimensional chaotic maps (4D, 3D, 2D and 1D etc.). Each P-value determines whether the produced sequence is random in nature or not. A P-value equals to 1 determines perfect randomness. If P is in the range of 0.01 to 1, then the test indicates that the sequence produced is completely random in nature. The randomness of the sequence generated by the proposed algorithm can be evaluated by converting the encrypted pixels P_i to bit P_{ib}. The NIST Table 9 shows that it is successful against statistical attacks and hence the proposed method is feasible for cryptography applications.

13. CONCLUSION

The proposed work classifies the sample document images and assigns them a priority value automatically based on the type of the document along with its features and encrypts each document with different levels of security. The performance of the proposed method is evaluated by subjecting different types of sample Document images to the classifier and then to encryption. With more number of features of the image and a few neighbors enabling the classification of the image correctly and efficiently. The magnitude of security is depicted in Table 6 for different document classes with the parameters, the Entropy, the Mean Square Error, PSNR, Correlation, Variance, NPCR, UACI, SSIM, and UIQ. The results are obtained for a bunch of documents of different types. The documents are perfectly classified and correct priorities are assigned. The documents encrypted with highest priority have highest noise, randomness, ideal pixel change rate and ideal unified average change intensity with better correlation among the neighboring pixels when compared to the documents encrypted with lower priorities. The Table 6 reveals that the proposed method is equipped to resist statistical and differential attacks with variable security levels. It is found that encryption of a set of Document images without a priority results in more encryption time when compared to the documents with priority as in Table 4. Thus encryption of images followed by the classification with priorities is saving the system resources and also providing the required security against the attacks. The NIST test is to check for the randomness in the chaotic sequence yielded complete randomness as depicted in Table 9. The use of different dimensional chaotic maps also contributed for the variable security levels and encryption time. Based on the priority value, a different second level diffusion technique uses a complex method for establishing interdependency between pixels involves more mathematical operations. The cipher images obtained for different input document images are different from one another and there is non-linearity in them. This makes the crypt analyzer difficult to decrypt the images with proper key to extract the original image. Hence the proposed method encrypts different types of documents with variable security levels and encryption time.

References

[1] S. Marinai, Introduction to Document Analysis and Recognition. Springer-Verlag Berlin Heidelberg, 90, pp. 1-20, 2008.

[2] A. Kumar, F. Nette, K. Klein, M. Fulham and J. Kim, “A visual Analytics Approach using the Exploration of Multi-Dimensional Feature Spaces fo Content- based Medical Image Retrieval”, *IEEE Journal of Biomedical and Healt Informatics*, pp. 168-2194, 2013.

[3] N. Chen, D. Blostein, A survey of Document image Classification: Problem Statement, Classifier architecture and Performance Evaluation. Springer-IDJAR, 2006.

[4] O. Augereaw, N. Journet, J-P. Domenger, “Document images Indexing with Relevance Feedback: an Application to Industrial Context”, In Proceedings of the International Conference on Document Analysis and Recognition, IEEE Computer Society, 2011, pp. 1190-1194.

[5] F. Chen, A. Girgensohn, M. Cooper, Y. Lu and G. Filby, “Genre Identification for Office Document search and browsing. Springer-IDJAR”, 2012, pp. 167-182.

[6] S. Sergyan, “Color Histogram Features Based Image Classification in Content- Based Image Retrieval Systems”, In Proceedings of the 6th International Symposium on Applied Machine Intelligence and Informatics, 2008.

[7] F. Esposito, D. Malerba and F. A. Lisi, “Machine Learning for Intelligent Processing of Printed Documents”, *Journal of Intelligent Information Systems*, vol. 14, pp. 175-198, 2000.

[8] V. Eglin, S. Bres, L.-Rfv, I. de Lyon, “Document page Similarity based on layout visual saliency: Application to query by example a Document Classification”, In Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR-2003), IEEE-Computer Society, 2003.

[9] A. Schenker, M. Last, H. Bunke and A. Kandel., “Classification of Web Documents using a Graph model”, In Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR-2003), IEEE-Computer Society, 2003.

[10] E. Appiani, F. Cesarini, A.M. Colla, M. Diligenti, M. Gori, S. Marinai and G. Soda, “Automatic Document Classification and Indexing in high-volume Applications”, Springer-Verlag (IJDAR), pp. 69-83, 2001.

[11] Ms. K. Arthi and Mr. J. Vijayaraghavan, "Content based Image Retrieval Algorithm using Color Models", *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, Issue 3, 2013.

[12] S. Shastry, G. Gunasheela, T. Dutt, D. S. Vinay and S. R. Rupanagudi, ""i"-A novel algorithm for optical character Recognition (OCR)", IEEE, pp. 389-393, 2013.

[13] A. Farhat, A. Al-Zawqari, A. Al-Qahatni, O. Hommos, F. Bensaali, A. Amira and X. Zhai, "OCR Based Feature Extraction and Template Matching Algorithms for Qatari Number Plate", IEEE, 2016.

[14] C. R. Revanna and Dr. C Keshavamurthy, "A Secure Document Image Encryption Using Mixed Chaotic System" *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 15, no. 3, pp. 263-270, 2017.

[15] C. R. Revanna and Dr. C Keshavamurthy, "A New Selective Document Image Encryption Using GMM-EM and Mixed Chaotic System", *International Journal of Applied Engineering Research*, vol. 12, pp. 8854-8865, 2017.

[16] H. Liu, and X. Wang, "Color image encryption using spatial bit-level permutation and High-dimension chaotic system"', *Optical Communication,* vol. 284, pp. 3895–3903, 2011.

[17] A. Melo, P. Bezerra, and A. Ablem, et al. "Priority QoE: a tool for Improving the QoE in Video Streaming", Intelligent Multimedia Technologies for Networking Applications: Techniques and Tools, chapter 11.

[18] N. Shaikh, S. Chapaneri and D. Jayaswal, "Hyper Chaotic Color Image Cryptosystem", In Proceedings of the IEEE International conference on Advances in Computer Application, 2016, pp. 239-243.

[19] V. Praneeth Kumar Reddy and A. Annis Fathima "A cost Effective Approach for Securing Medical X-ray images using Chebyshev Map", In Proceedings of the IEEE 5th International Conference on Recent Trends in Information Technology, 2016.

[20] M. Dridi, M. Ali Hajjaji, B. Bouallegue and A. Mtibaa, "Cryptography of medical images based on a combination between chaotic and neural Network", *IET, Image Processing*, pp. 1-10, 2016.

[21] B. Awdun and G. Li, "The Color Image Encryption Technology based on DNA Encoding and Sine Chaos", In Proceedings of the IEEE International conference on Smart City and System Engineering. 539-544, 2016.

[22] IEEE Computer Society. (1985). IEEE standard for binary Floating-Point Arithmetic, ANSI/IEEE standard, August 1985, p. 754.

[23] G. Alvarez, and S.J. Li, "Some basic cryptographic requirements for chaos-based Cryptosystem", *Int. J. Bifurcation Chaos*, vol. 16, no. 8, pp. 2129-2151, 2006.

[24] G. Ye, "A block image encryption algorithm based on wave transmission and chaotic systems", Springer-Nonlinear Dyn, vol. 75, pp. 417-427, 2014.

[25] X. Wang, L. Liu, Y. Zhang, "A novel chaotic block image encryption algorithm based on dynamic random growth technique. ELSEVIER Optics and laser in Engineering, vol. 66, pp. 10-18, 2015.

[26] Z. Yu, Z. Z. Yang et al., "A Chaos-Based Image Encryption Algorithm Using Wavelet Transform", In Proceedings of the IEEE Conference, pp. 217-222, 2010.

[27] D. E. Goumidi, F. Hachouf, "Hybrid chaos based image encryption approach using block and stream ciphers", In Proceedings of the IEEE international workshop on system signal processing and their applications, 2013, pp.139-144.

[28] S. M. Wadi and N. Zainal, High Definition Image Encryption Algorithm Based on AES Modification. Springer Science Business Media New York, pp. 811-829, 2014.

[29] Y. Zhang, X. Li and W. Hou, "A Fast Image Encryption Scheme Based on AES", In Proceedings of the 2nd International Conference on Image, Vision and Computing, 2017, pp. 624-628.

[30] Y. Zhang, "Test and Verification of AES Used for Image Encryption", Springer-Verlag GmbH Germany, 2018.