Improving Accuracy of CADx System by Hybrid PCA and Backpropagation

Hind S. Harba
Department of Atmospheric Sciences, College of Science,
University of Mustansiriyah
Baghdad, Iraq
e-mail: hindharba76@yahoo.com

Eman S. Harba
Computer and Internet Unit, College of Arts
University of Baghdad
Baghdad, Iraq
e-mail: emanharba_1212@yahoo.com

Samera Shams Hussein
Department of Computer Science, College of Education
for pure Science, University of Baghdad
Baghdad, Iraq
e-mail: sameramazn@yahoo.com

Mohammed Kahdoum Farttoos
Directorate of Nuclear Application
Ministry of Sciences and Technology
Baghdad, Iraq
e-mail: mkf.iraq@gmail.com

***Abstract*—Medical images have recently played a significant role in the diagnosis and detection of various diseases. Medical imaging can provide a means of direct visualization to observe through the human body and notice the small anatomical change and biological processes associated by different biological and physical parameters. To achieve a more accurate and reliable diagnosis, nowadays, varieties of computer aided detection (CAD) and computer-aided diagnosis (CADx) approaches have been established to help interpretation of the medical images. The CAD has become among the many major research subjects in diagnostic radiology and medical imaging. In this work we study the improvement in accuracy of detection of CAD system when combined principal component analysis and feed forward back propagation neural network. This work has investigated the ability to improve the CAD system in order to use in detection abnormality even with low cost diagnosis methods (such as mammogram images or X-ray). The results show that the reduction of correlated details within the training data by using the PCA method can enhance the recognition performance. The performance of the neural network diagnostic to discriminate the normal cases from cancerous cases, evaluated by using recognition analysis show a high accuracy in detection. The proposed approach can be considered as a potential tool for diagnosis breast cancer from x-ray and mammography images and prediction for non-experts and clinicians.**

Keywords-principal component analysis; feed forward back propagation neural network; breast cancer; computer-aided diagnosis; medical imaging

I. INTRODUCTION

Medical images are a significant part of medical diagnosis and thus for the treatment we concentrate on these images for enhanced results. The medical images involve features and hidden information that is used by the doctors to make relevant decisions about the situation of the patient. Hence, there is a need to utilize a data mining technique for detecting abnormality from medical images. X-ray imaging is the one of the most economical and effective imaging modality which are commonly used in medicine; and the mammogram the second most commonly used type of diagnosis of breast cancer tests. Both used widely in detecting breast tumor, because of low cost, simplicity, and effectiveness. Thus, the improving a method to enhance detection of abnormality in the mammogram and x-ray images are the very important to help the radiologists in the diagnosis and determine the patient case [1].

Modern studies are focused on utilizing as an aided tool for medical diagnosis, which is known as computer-aided diagnosis (CAD) that represent a diagnosis that made by a physician who considers the computer output as a second point of view. The CAD purpose is to enhance the diagnostic accuracy and the consistency the of the radiologist's image interpretation. Nowadays, CAD has become a part of the routine clinical work for detection of different cases such as detect the breast cancer from mammograms, lung tumor from x-ray images, etc. [2].

The general structure of CAD system usually includes three basic components that based upon three different technologies. The first part is image processing, which used to enhance and extract of lesions. It is very important to CAD to be involved image processing because it can be facilitating the computer, instead of humans observing, to determine the suspicious patterns and the initial candidates of lesions. The second part is the image features quantitation such as the contrast, size, and the shape of the candidates selected in the first step. It's possible to determine different features depends upon some mathematical formula which may not be readily understood by the human observer. Even so, it is beneficial to define the features of image that have previously been recognized and described by radiologists [3]

The third part is the data processing for distinction between abnormal and normal patterns, depending on the features obtained in previous part. A common and simple approach utilized in this part is a rule-based method that can be established depending on the comprehension of lesions and other normal patterns. Thus, it is necessary to note that the rule-based approach can provide beneficial information

978-1-5386-6995-2/18/$31.00 ©2018 IEEE

for improving the CAD. The other techniques utilized include artificial neural network (ANN), decision-tree method, and the discriminant analysis. The artificial neural network is one of the most efficient method that can utilized in that purpose, that can produce the best results in CAD [3].

The applications of ANN in CADx represent the main field of computational intelligence in medical imaging. Their involvement and penetration is most inclusive for all medical problems because the neural networks have the adaptive learning nature of input data and, utilizing a proper learning algorithm, can upgrade themselves in accordance with the varieties and the change of input content. Moreover, ANN has the ability of enhancing the relationship between the outputs and inputs through training, processing, and distributed computing, which leads to reliable solutions eligible by specifications, and medical diagnosis [4].

Principal component analysis is a statistical method for converting potentially correlated variables of observation data into a set of linearly uncorrelated variables called principal components. PCA technique transforms an image data into small valued data which can be easily handled by neural network, which can reduce the dimension of the input space in the ANN models, while decreasing the time of training and preserving or even enhancing the ANN model accuracy [5].

The aim of this paper is to investigate the effects of PCA in improving BPNN for CADs applications of detecting normal and cancerous cases from mammograms and x-ray images.

II. System Methodology

In this work, a hybrid technique has been proposed in which utilize hybridization of PCA and Back Propagation Neural Network (BPNN). Here, first a Principal Component Analysis has been applied to the feature extraction of the images obtained for detection. We used Backpropagation Neural network to train obtained features after applying the PCA algorithm and once the features are trained, then compared the results of hybrid method with normal training method to compute the accuracy gain.

A. *Feature Extraction*

Principal Component Analysis (PCA) has been applied for feature extraction. PCA is basically beneficial for minimizing the number of variables which consists a dataset whilst retaining the inconsistency in the data and to determine the unknown patterns within the data and to classifies them based on how much information stored in the data. PCA permits the calculating the linear alteration that maps information from high dimensional space to a lower dimensional space.

B. *Architecture of Artificial Neural Network*

The architecture of ANN (shown in Fig. 1) is a multi-layer neural network that's been utilized for pattern recognition. In this type of ANN, the input from the high number of input layer is fed to a lower number of neurons in the hidden layer, then is further fed to the higher number of neurons in the output layer. This type of network is also known as bottleneck feed forward neural network (FFNN). The multilayer back propagation neural network is one of the most important types of FFNN [6].

The neural network architecture for image compression, for example 8x8 image (Fig.1) that contain 64 input neurons, 64 output neurons and 16 hidden neurons based on the requirements.

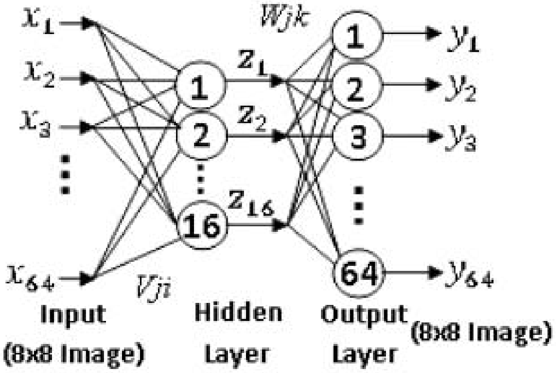


Figure 1. Bottleneck type FFNN. [6]

The input layer encodes the neuron inputs and moves the encoded information toward the hidden layer neurons. The output layer receives the information of hidden layer and then decoding the information in the output. The hidden layer outputs are real valued and needs many bits to transfer the data. The input layer (transmitter) encodes and then transmits 64 values of the original image to 16 values of the hidden layer. Then from hidden it will transfer to output layer (the receiver), which is receives and decodes the 16 hidden neuron's output and generates the 64 outputs at the output layer. The techniques like PCA are needed to incorporate with ANN to manage the output and input image data [7].

III. Proposed System Architecture

The stepwise procedure of the proposed work is as follows:

Step 1: Firstly, it is required to collect the various images to make a database of these images. The processing method, which consists of finding the edges of the images and their normalization, is applied. It makes the further proceedings easy.

Step 2: Principal Component Analysis will then be applied to the feature extraction of the images.

Step 3: The neural network is trained by using the data extracted from the images. The breast cancer will be classified by using neural network.

Step 4: Performance parameters like Accuracy, False Rejection Ratio, and False Acceptance Ratio, and will be evaluated.

Fig. 2 illustrated the depicting the proposed system methodology.

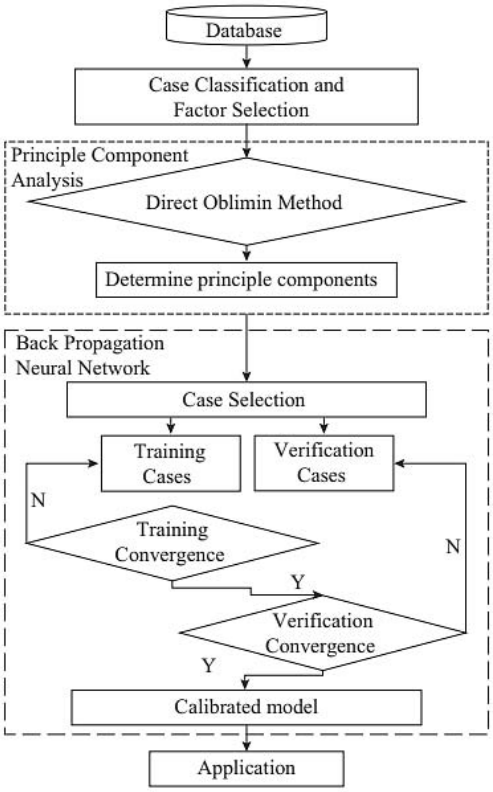


Figure 2. Proposed system methodology architecture

A. Database

A mammograms medical image technique has been selected in order to identify the accuracy of the system. The mammography images got from Digital Database for Screening Mammography (DDSM) for two cases (normal and cancer). A data base has been built for and 60 images taken for each case from the both (60 normal, 60 benign and 60 cancer) images from DDSM and (60 normal and 60 cancer) images from x-ray center.

B. Pre-processing of Images

After acquisition of monograph image, each of these images have gone little pre-processing procedures, including normalization, grayscale conversion, etc.

1. **Grayscale Conversion**
   This method requires conversion of original mammogram images that are in RGB format to grayscale by eliminating saturation and tint information while keeping the luminance.
2. **Normalization**
   After grayscale the mammogram images, then converted it from its original size to (64x64) pixels' values. This specific size provides reliable information in low processing time.

C. Feature Extraction

PCA is generally ideal for lowering the number of factors that comprises a dataset whilst keeping the inconsistency within the data and to recognize unknown patterns from the data as well as categorize them based on the amount of the information, saved in the data. PCA enables determining a linear alteration of which maps information by a higher dimensional space into a lower dimensional space [8].

D. Neural Network Classifier

A standard, 3-layer BPNN has been utilized in Intelligent System for detection breast tumor with 64x64 input images, that have 4096 input neurons, and used 2700 hidden neurons and two output neurons that categorize the breast to breast with tumors and with no tumors. The mammography images are sorted in binary coding as [1 0] cancerous breast and [0 1] for the normal breast. The activation function (sigmoid) has been used for activating purpose of the neurons both in output and hidden layers. The BP network is illustrated in Fig. 1. Due to the simplicity of implementation, as well as the availability of an adequate database (input target") for training, utilizing a BPNN that is a supervised learner, is often preferred. Generalization (and Training testing are made in this stage.

The backpropagation algorithm can be summarized as follows:

Step 1: Determine the network architecture

- Determine the input and output neurons; and output labels
- Determine hidden neurons and layers

Step 2: Initialized the activation of the neural network. The thresholding unit's values should not change. a) X = 1.0 b) h = 1.0

Step 3: Select an input and output pair. Assume the input vector is Xi and assume the Yi is the target output vector. Give activation levels for the input units.

Step 4: Propagate the activation function from input units to the hidden units by utilizing the activation functions [8]:

$$\Delta h_j = \frac{1}{1+e^{-\sum_{i=0}^{n} W1_{ij}}}, \text{ For j = 1, ..., n} \qquad (1)$$

The i ranges from 0 to B. $W1_{ij}$ is the thresholding weight for j.

Step 5: Propagate the activation function from the input units to the hidden units using the activation functions [8]:

$$\Delta h_j = \frac{1}{1+e^{-\sum_{i=0}^{n} W2_{ij}}} \qquad (2)$$

For j=1, …., n

The thresholding weight $w2_{ij}$for output unit j is important in the weighted summation. where h is 1.

Step 6: Calculate the units' errors inside the output layer denoted $\delta 2_j$. The Error is dependent on the network real output (o_j) also, the target output (Yi) [8]:

$$\delta 2_j = o_j(1 - o_j)(y - o_j) \qquad (3)$$

For j = 1, …., n.

Step 7: Calculate the units errors in the hidden layer, denoted $\delta 1_j$ [8]:

$$\Delta 1_j = h_j(1 - h_j) \sum_i^c \delta 2_i \times w_{ij} \qquad (4)$$

For j = 1, …., n, i= 1, …., m

$$\Delta w1_{ij} = \eta \times \delta 2_j \times h_j \qquad (5)$$

For all i = 0, …., m, j=1…., n

Step 8: Modify the weights between hidden and output layer. The learning rate denoted represent the denoted by η; and it is functions are similar to perceptual learning. A sensible value of η is 0.35 [8].

$$\Delta w2_{ij} = \eta \times \delta 2_j \times h_j \qquad (6)$$

For all i = 0,…., n, j=1…,m

Step 9: Modify the weights between input and hidden layer [8].

$$\Delta w1_{ij} = \eta \times \delta 1_j \times h_i \qquad (7)$$

For all i = 0, …., A, j=1….,B

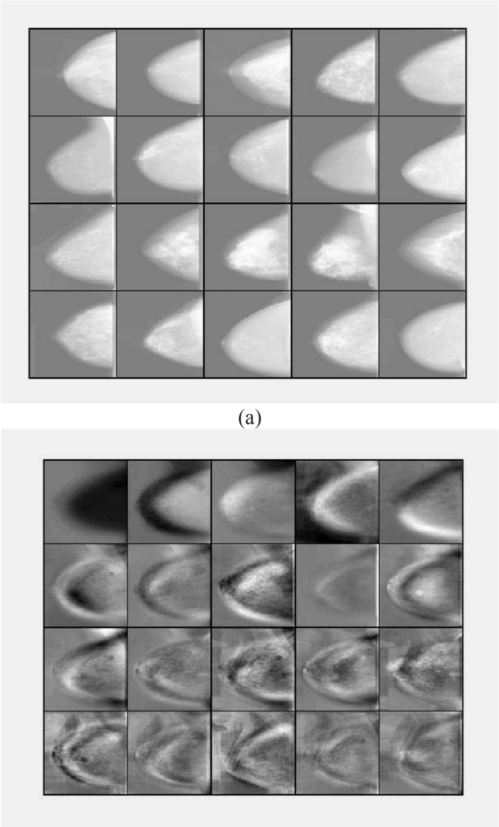
Step 10: Return to step 3 and do the steps again. As soon as all of the inputs/output pairs have been completely given to the network, a one epoch has become accomplished. Also, repeating steps from 3 to 10 based on how many epochs is desired.

The procedure for this proposed procedure can be summarized as follows:

1. **Step 1:** collecting various breast cancer images to make a database of these images.
2. **Step 2:** The pre-processing method, which consists of converting images to grayscale and normalizing it to desired scale, which makes the further proceedings easy.
3. **Step 3:** Principal Component Analysis will then be applied to the feature extraction of the images. Features will be optimized by using Bacterial Forging Optimization.
4. **Step 4:** The neural network is trained by using the data extracted from the images. The breast cancer will be classified by using neural network.
5. **Step 5:** Performance parameters such as: Accuracy, False Rejection Ratio and False Acceptance Ratio has been evaluated.

IV. Results and Discussion

For first we run PCA and visualize the vectors, then visualization of mammogram images after PCA dimension reduction. The result of applying PCA reduction is shown in Fig. 3.



(a)

(b)

Figure 3. Processing data: (a) original images, (b) visualization of mammogram images after pca dimension reduction

We have been trained the PCA data by using BPNN learning phase used initial random weights that have a value in a range between -1 and 1. So as to achieve requested minimal error value we selected iteration to be 100. The learning result has been shown in Fig. 4, which represented the improvement of recognition (error reduction) per iteration.

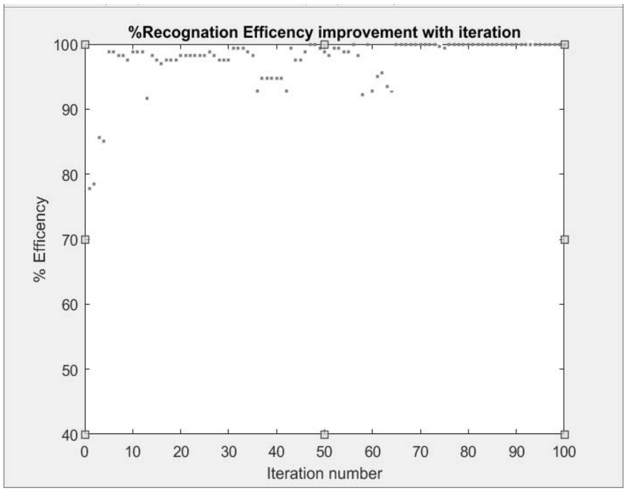


Figure 4. Recognition efficiency per iteration, x-axis the efficiency and y-axis the iteration

As shown from Fig. 4, the efficiency of recognition has been increased with iteration as a result of reduction in cost function and it has been reached to stability after 70 iterations to be ~100%. This higher value proves the improvement in the training process by PCA. Fig. 5 shows side by side recognition result of sampled of mammogram images.

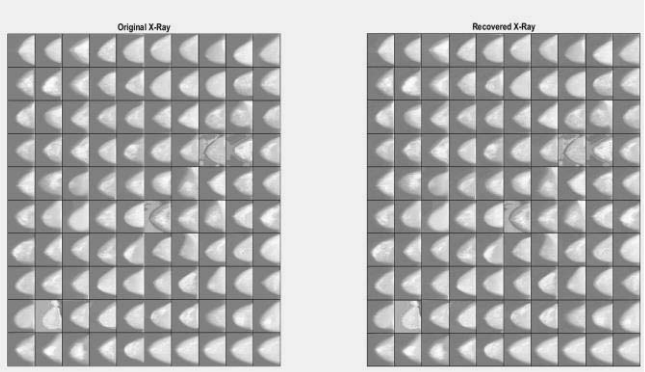
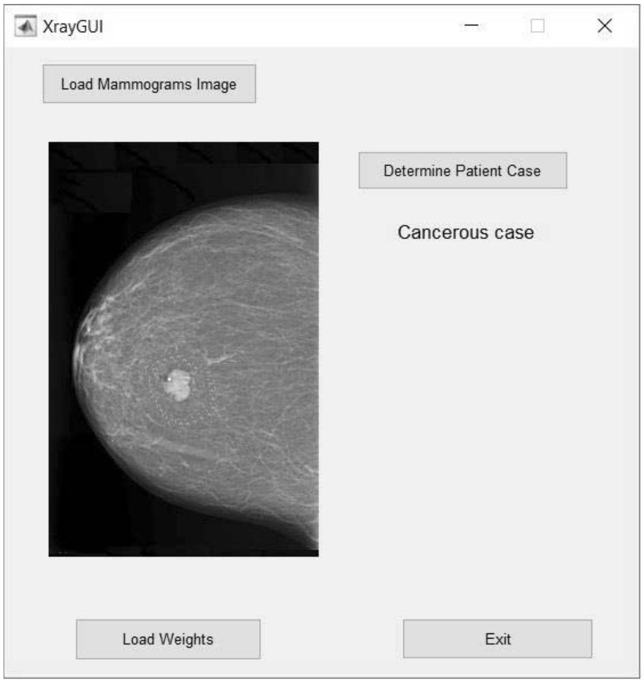


Figure 5. Comparison between the original x-ray images (left) and the recognized image (right)

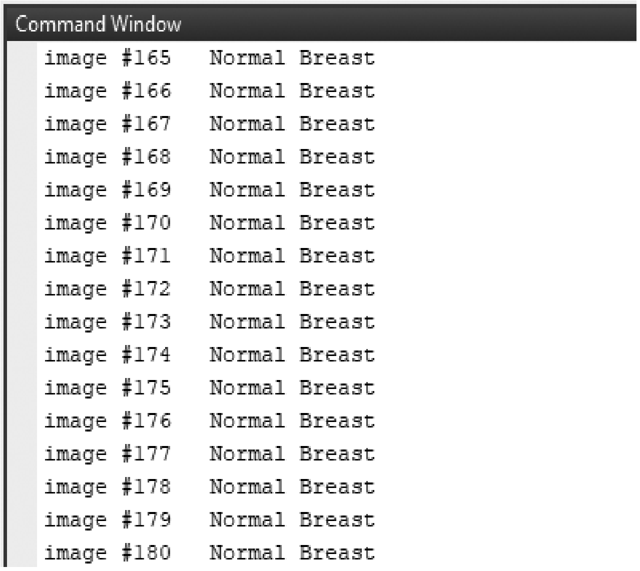
As shown in Fig. 5, where the image at left is original trained images and the right is recognized output image that approved the recognition has a 100% detection.

To verify the real recognition test we test the recognition for trained and non-trained images that had checked by special radiologist to determine the best case. The result shows a high accuracy in recognition and detection breast that have tumors or a normal one. Fig. 6 shows the diagnosis result for non-trained x-ray images and Table I shows the recognition results of trained and non-trained x-ray images.



Cancerous case

(a)



```
image #165   Normal Breast
image #166   Normal Breast
image #167   Normal Breast
image #168   Normal Breast
image #169   Normal Breast
image #170   Normal Breast
image #171   Normal Breast
image #172   Normal Breast
image #173   Normal Breast
image #174   Normal Breast
image #175   Normal Breast
image #176   Normal Breast
image #177   Normal Breast
image #178   Normal Breast
image #179   Normal Breast
image #180   Normal Breast
```

(b)

Figure 6. Recognition result for (a) normal case of trained image folder, (b) GUI window when detect the nontrained cancer case

The recognition result for trained and untrained images of normal and cancerous cases are shown in Table I.

TABLE I. RECOGNITION RESULTS FOR TRAINED AND UNTRAINED IMAGES

Method	Mammography Images Cases			
	Cancer (180 trained, 30 untrained)		Normal (180 trained, 30 untrained)	
	Detect	Failed	Detect	Failed
Trained	176	4	179	1
Untrained	25	5	27	3

As shown from Table I, the recognition of trained images achieved higher accuracy (97.77%) in the detection of cancerous image and (99.4%) for detecting normal cases. The recognition results of nontrained images achieved good accuracy for untrained images, which is (83.33%) in the detection of cancerous image and (90%) for detecting normal cases.

V. CONCLUSION

In this paper, a CADx system based on used (PCA) as a feature extraction method and backpropagation neural network (BPNN) has been presented. The proposed system has been utilized to diagnose the breast. The PCA has been used to reduce the dimension as well as improved the BP-ANN performance in terms of execution time. Successful system implementation, has been implemented, to identify the normal Brest and breast that have tumors. Considering known and unknown cased that represented by non-trained images, the accurate recognition results got (97.77%) in the detection of cancerous image and (99.4%) for detection of normal cases. Since untrained images are (83.33%) in the detection of cancerous image and (90%) for detecting normal cases. This assists the doctors and radiologists for easy distinction between normal and cancerous breast. From our study, it provides that the using (PCA) with BPNN is more accurate than using it alone in detection of breast cancer.

REFERENCES

[1] Jadhav, A.; D'Cruz, J.; Chavan, V.; Dighe, A.; Chaudhari, J. Detection of Lung Cancer Using Backpropagation Neural Networks and Genetic Algorithm. International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4 , April 2016.

[2] Castellino, R. A. Computer aided detection (CAD): an overview. Cancer Imaging. 5(1), 2005, pp.17–19.

[3] Doi, K.; MacMahon, H; Katsuragawa, S.; Nishikawa, R. M.; and Jiang, Y. Computer-aided diagnosis in radiology: potential and pitfalls. Elsevier Science Ireland Ltd., European Journal of Radiology, 1997, pp.97-109.

[4] Jiang, J .; Trundle, P.; and Ren, J. Medical Image Analysis with Artificial Neural Networks. Digital Media & Systems Research Institute, University of Bradford.

[5] Jolliffe, I.T.; Cadima, J. Principal component analysis: a review and recent developments. Royal Society, Philos Trans A Math Phys Eng Sci. 3; 374(2065), 2016, doi:10.1098/rsta.2015.0202.

[6] Kurkova,V.; Steele, N.C.; Neruda, R.; Karny M. Artificial Neural Nets and Genetic Algorithms: Proceedings of the International Conference in Prague, Czech Republic. Springer Science & Business Media, 2013, ISBN: 9783709162309.

[7] Gaidhane,V.H.; Singh, V.; Hote,Y.V.; Kumar, M. New Approaches for Image Compression UsingNeural Network. SciResJournal of Intelligent Learning Systems and Applications, Vol. 3, 2011, pp.220-229, doi:10.4236/jilsa.2011.34025

[8] Satapathy, S.C.; Udgata,S.K.; Biswal, B. N. Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA). Springer Science & Business Media, ISBN: 9783319029313.