# Image Spam Classification Using Neural Network

Mozammel Chowdhury[1(✉)], Junbin Gao[1], and Morshed Chowdhury[2]

[1] School of Computing and Mathematics, Charles Sturt University, Bathurst, Australia
{mochowdhury,jbgao}@csu.edu.au
[2] School of Information Technology, Deakin University, Geelong, Australia
muc@deakin.edu.au

**Abstract.** Spam, an unsolicited or unwanted email, has traditionally been and continues to be one of the most challenging problems for cyber security. Image-based spam or image spam is a recent trick developed by the spammers which embeds malicious image with the text message in a binary format. Spammers use image based spamming with the intention of escaping the text based spam filters. On the way to detect image spam, several techniques have been developed. However, these techniques are vulnerable to most recent image spam and exhibit lack of competence. With a view to diminish the limitations of the existing solutions, this paper proposes a robust and efficient approach for image spam detection using machine learning algorithm. Our proposed system analyzes the file features together with the visual features of the embedded image. These features are used to train a classifier based on back propagation neural networks to detect the email as spam or legitimate one. Experimental evaluation demonstrates the effectiveness of the proposed system comparable to the existing models for image spam classification.

**Keywords:** Image spam · Spam filtering · Machine learning · BPNN

## 1 Introduction

Nowadays, e-mails have become a very common and convenient medium to millions of people worldwide for daily communications due to the rapid advances of Internet. However, along with the emergent significance of the emails, there has been a striking growth of spam in recent years which has become a key problem to the internet users and vendors. Spam is commonly defined as an unsolicited or unwanted bulk e-mail sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipients [1]. The current trend of spam messages alarms that it will climb to 95% of the total email traffic very shortly, which was accounted about 70% in 2012 [2]. Due to the recent upsurge in spam emails, it has been a significant concern for the researcher to develop unbeaten techniques for fighting against spam.

Until last decade, the spam messages were based on textual content only. That's why, the spam filters [3-6] were designed to analyze only the text content of the messages to classify them as spam or legitimate email. However, in recent years, spammers has introduced a new trick by developing multimedia enriched spam, where the

B. Thuraisingham et al. (Eds.): SecureComm 2015, LNICST 164, pp. 622–632, 2015.
DOI: 10.1007/978-3-319-28865-9_41

text message is embedded into the attached image with an intention to defeat the text-based anti-spam filters. Fig. 1 shows the examples of spam images. Approaching to detect and filter image spam, several techniques have been recently proposed [7-12]. However, these proposed solutions exhibit several weaknesses and their effectiveness has not been thoroughly investigated so far.

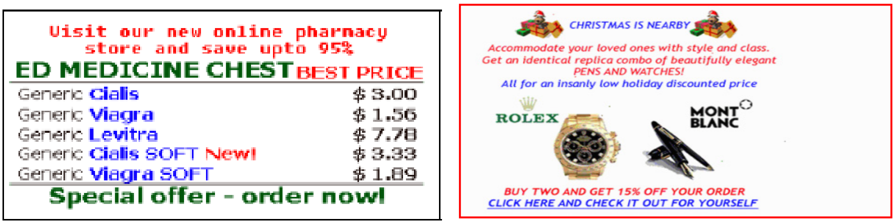

Visit our new online pharnacy
store and save upto 95%
ED MEDICINE CHEST BEST PRICE
Generic Cialis $ 3.00
Generic Viagra $ 1.56
Generic Levitra $ 7.78
Generic Cialis SOFT New! $ 3.33
Generic Viagra SOFT $ 1.89
Special offer - order now!
CHRISTMAS IS NEARBY
Accommodate your loved ones with style and class.
Get an identical replica combo of beautifully elegant
PENS AND WATCHES!
All for an insanly low holiday discounted price
ROLEX
MONT BLANC
BUY TWO AND GET 15% OFF YOUR ORDER
CLICK HERE AND CHECK IT OUT FOR YOURSELF

**Fig. 1.** Examples of spam images: (a) image with embedded text (b) image with text and picture.

Many researchers have contributed to fight against the arms racing of spam by developing new techniques. In recent years, machine learning based text categorization techniques have been widely investigated for textual content analysis [13-17]. The success of machine learning techniques for text categorization has inspired researchers to explore learning algorithms in developing spam filtering. In particular, Bayesian techniques and Support Vector Machines (SVM) are most effective methods for text categorization, which are widely used by the researchers for spam classification [3].

It is a matter of fact that the unbeaten response of the content-based filters has forced spammers to originate increasingly complex attacks to escape these filters. On the way to struggle against the spammers' tricks, researchers have employed learning capability with these filters to train those using machine learning algorithms. Learning-based filters have the potential to learn and enhance the self-performance at real-time, so that they can adapt themselves to the wide genre of spam.

In this paper, a new architecture of spam classification has been proposed based on back propagation neural network (BPNN). The system will analyze the file features of the embedded image and extract the low level visual features as well. These features are then fed into the BPNN classifier to train the network. To test the effectiveness of the proposed network and verify the accuracy, we use a large data set consisted of both spam and non-spam images. Experimental evaluation confirms that the proposed system is robust and efficient to detect the embedded message as spam or legitimate email.

The remainder of the paper is organized as follows. Section 2 provides an overview of relevant work in this research area. Section 3 describes our proposed approach for image spam classification. Section 4 demonstrates the experimental results and performance of the proposed system with a critical discussion. Finally, Section 5 concludes the paper with future research directions.

## 2 Related Works

Many techniques have been proposed by the researchers in last recent years for detecting image spam. In this section we provide a brief discussion on relevant work in image spam classification.

Wu *et al.* [18] proposed an image spam classification technique based on text area and low-level features of the image. They argued that computer-generated graphics like banner, advertisement are spam images attached with emails. They considered the ratio of the banners and graphic images to the total number of attached images as features based on the assumption that most of the spam images are banners and computer-generated graphics as advertisements. Banners were detected considering the aspect ratio, height, and width. To identify the computer-generated graphics they assumed that graphics contain homogeneous background and less texture. A one class classifier based on SVM was used in their work.

Aradhye *et al.* [19] proposed a technique for image spam detection based on extracted overlay text and color features. It can monitor outbound e-mails by corporations to detect communications including proprietary or confidential material of the corporation. The method consists of three stages: (i) extraction of the text containing in the spam image, (ii) identification of spam-indicative features from the image, and (iii) learning the features with a SVM for image spam categorization.

A fast classifier using Maximum entropy, Naïve Bayes and Decision tree was proposed by Dredze *et al.* [20] based on image metadata and low-level features. The technique exploits information like image height, width, aspect ratio, file format (e.g., gif, jpg), and file size. Visual features like average red, green and blue values, features based on edge detection were also considered.

Wang *et al.* [21] proposed an image spam classification technique based on low-level features and similarity of images. The similarity measure is estimated for each set of features. The distance measure is then compared to a threshold. The threshold is set different for each feature space. Based on the threshold value, the image is detected as spam or legitimate one. The image features are extracted from color histograms, Haar wavelet transform, and edge orientation histograms. They used Nearest neighbour detection in their technique.

Another image spam classification algorithm based on low level image processing technique was proposed by Biggio *et al.* [22]. This method can recognize the noisy texts in the malicious image. This technique can identify the presence or absence of noisy text, or measure the amount of noise in a proper scale.

Mehta *et al.* [8] proposed a two-class SVM classifier based on the low level color features and similarity of images. Their proposal assumed that spam images are artificially generated and are related to color, shape and texture of the images. Their distribution was approximated with Gaussian mixture models. They stated that the low-level features could help the email recipients to achieve the highest capability for discriminating the spam and non-spam images.

Zhang *et al.* [23] proposed a technique based on image similarity where similarity is computed on the basis of color, texture, and shape features of the image. They used a two-class SVM classifier trained on spam and legitimate images. This technique consists of three steps: (i) image segmentation, (ii) feature extraction and similarity calculation and (iii) spam image clustering.

Bowling *et al.* in [24] suggested an approach for image spam classification using artificial neural networks. Their method identifies image spam by training an artificial neural network. The process consists of three steps. The initial step is the image preparation. In the next step the neural network is trained with training data. In the final stage, the neural network is tested to identify whether the embedded image is spam or non-spam. The neural network was implemented with 22,500 inputs, two hidden layers of 50 or 75 nodes each, and one output node. The input nodes are the pixels of an image. The output layer is the +1 or -1 indicating spam or non-spam.

## 3 Proposed Architecture of Image Spam Detection Technique

The overall framework of our proposed method for image spam detection is shown in Fig. 2. The aim of this paper is to develop a classifier that can detect the image spam and legitimate emails. The proposed system consists of three main components: (i) Features extraction, (ii) Features selection and (iii) BPNN Classification. This section presents the proposed methodology for extracting the feature points from the embedded image and a feed forward back propagation neural network, which pretends as a classifier for detecting the image as spam or legitimate one.

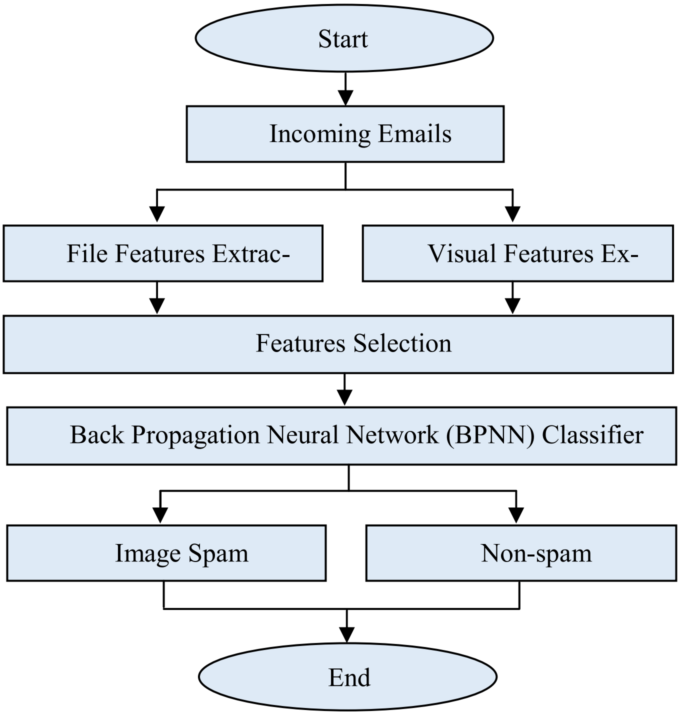


**Fig. 2.** Proposed approach for image spam classification.

### 3.1 Features Extraction and Selection

One of the key tasks underlying image spam classification is feature extraction. This paper extracts two types of features for image spam classification: one is file features and another is visual or color features of the image. Selected features are then feed forward to BPNN classifier.

#### 3.1.1 File Features Extraction

Image spam can be detected based on their file type. The authors [11] derive some features of the image file for detecting image spam using decision trees and support vector machine. In this work, we only extract the basic file features of an image with an intension of requiring low computation cost. The basic useful features of an image file include: image file type, file size and the dimension (width and height) denoted in the header of the image file. Empirically we find that image spam mostly contains images of GIF (graphics interchange format), PNG (portable network graphics) or JPEG (joint photographic experts group) file types. Therefore, we consider these three image file formats in our work. The file features of an image are reported in Table 1.

**Table 1.** File features of an image

| File features | Description |
|---|---|
| $f_1$ | Image width denoted in header |
| $f_2$ | Image height denoted in header |
| $f_3$ | Aspect ratio: $f_1/f_2$ |
| $f_4$ | File size |
| $f_5$ | Image area: $f_1 \times f_2$ |
| $f_6$ | Compression: $f_5/f_4$ |

We can obtain the image dimensions by parsing the headers of the image files with a minimal parse. However, an issue related to GIF files is that there will be presence of virtual frames, which may be either larger or smaller than the actual image width [11]. This problem can be detected by decoding the image data. In addition to this problem, another issue could be impressed in case of corrupted images as well as PNG and JPEG images. This problem is that the lines near the bottom of the image will not be decoded properly and no further image data can be decoded after that point. This issue can be a useful trick to the spammers.

We measure the signal to noise ratio (SNR) to estimate the volume of information in the image obtained from the file features. The SNR can be defines as the following equation:

$$SNR = \left| \frac{\mu_{spam} - \mu_{leg}}{\sigma_{spam} + \sigma_{leg}} \right| \tag{1}$$

where, $\mu_{spam}$ is the mean value of the spam,

$\mu_{leg}$ is the mean value of legitimate or non-spam,

$\sigma_{spam}$ is the standard deviation of spam,

$\sigma_{leg}$ is the standard deviation of legitimate or non-spam.

The mean value of the binary features reflect the percentages of images in the respective formats. The feature f6 is the most informative feature beyond the binary image file that retains the amount of compression. The compression is better if more number of pixels is stored per byte.

### 3.1.2 Visual Features Extraction

The spammers usually design the spam messages using highly contrasting colors with an intention that the spam emails should be easily noticeable by the users [10, 18]. Based on this constraint, we use HSI (Hue, Saturation, and Intensity) color histogram to extract the visual features from the image. HSI color space is different from the RGB color space and it separates out the intensity from the color information. Intensity represents the value or brightness of a color, which is decoupled from the color information in the represented image. Fig. 3 shows a three dimensional representation of the HSI color space. The central vertical axis represents the intensity. Hue defines the angle relative to the red axis, and Saturation is the depth or purity of the color measured from the radical distance from the central axis with value between 0 at the center to 1 at the outer surface. This histogram is converted into three bins and passed into neural networks.

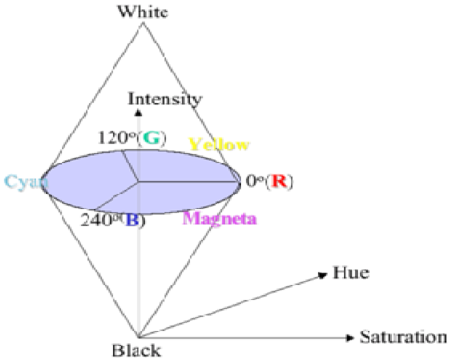


**Fig. 3.** Representation of HSI color space.

We can convert the RGB color space into HIS color space as follows:

$$I = \frac{1}{3}(R + G + B) \tag{2}$$

$$S = 1 - \frac{3}{R + G + B}\left[min\left(R, G, B\right)\right]$$

$$H = cos^{-1}\left[\frac{\frac{1}{2}\left[\left(R - G\right) + \left(R - B\right)\right]}{\sqrt{\left(R - G\right)^2 + \left(R - B\right)\left(G - B\right)}}\right]$$

### 3.2 The BPNN Classifier Model

A back-propagation neural network (BPNN) is a multi-layer artificial neural network consists of neurons [2, 24]. The layers are fully connected, that is, every neuron in each layer is connected to every other neuron in the adjacent forward layer and each connection has a weight associated with it. Back propagation algorithm presents a training sample to the neural network and compares the obtained output to the desired output of that sample. It calculates the error in each output neuron. BPNN adjusts the weights of each neuron to minimize the error. BPNN is used as a supervised training model for classification of image spam using the optimum feature vectors extorted from an image. It recognizes the data and test how well it has learned from the previous set of data.

Fig. 4 shows the architecture of the back-propagation neural network. The network consists of one input layer with 20 neurons and two hidden layers with 80 neurons and one output layer with a single neuron. The input nodes take the pixel values of an image. The output layer results -1 or 1 indicating non-spam or spam image respectively. The indices, *i, j, k*, refer to the neurons in the input, hidden and output layers, respectively. Input signals are propagated through the network from left to right, and error signals from right to left. The symbol $w_{ij}$ denotes the weight for the connection between neuron *i* in the input layer and neuron *j* in the hidden layer, and the symbol $w_{jk}$ the weight between neuron *j* in the hidden layer and neuron *k* in the output layer.

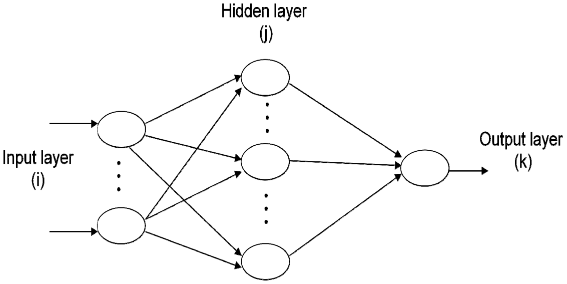


**Fig. 4.** The model of the back-propagation neural network

To propagate error signals, we start at the output layer and work backward to the hidden layer. The error signal at the output of neuron *k* at the $p^{th}$ training cycle (iteration) is given as:

$$e_k(p) = d_k(p) - y_k(p) \tag{3}$$

The instantaneous value of error energy for neuron k is,

$$E(i) = \frac{1}{2} e_k^2(p) \tag{4}$$

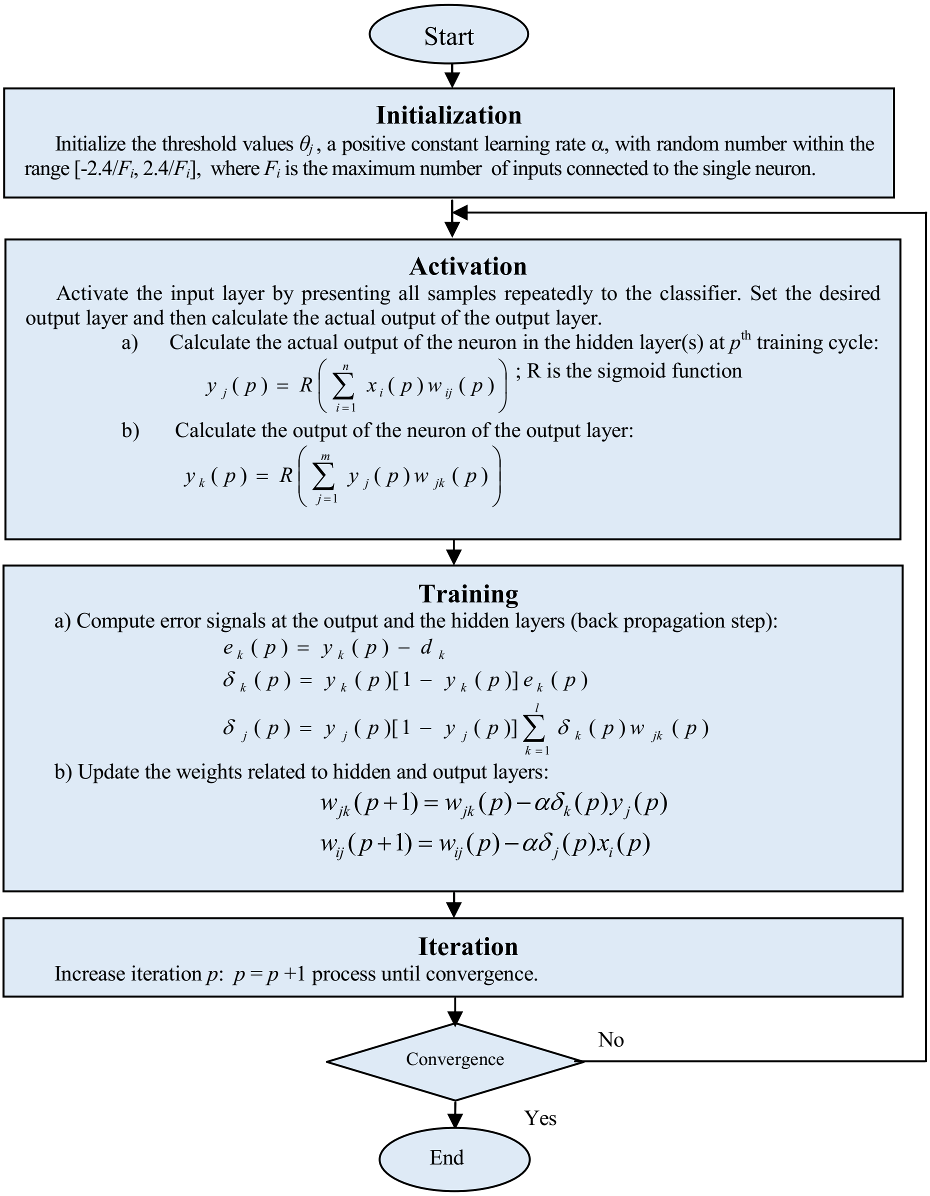
The total error energy *E*(p) can be computed by summing up the instantaneous energy over all the neurons in the output layer:

$$E(p) = \sum_{k \in C} \frac{1}{2} e_k^{\ 2}(p) \ ; \text{ C is a set of all output neurons} \tag{5}$$

The sigmoid function transforms the input, which can have any value between plus and minus infinity, into a reasonable value in the range between 0 and 1. The input value is passed through the sigmoid activation function. The sigmoid function can be expressed as,

$$R = \frac{1}{1 - e^{-x}} \tag{6}$$

Fig. 5 show the flow diagram of the BPNN classifier model.



**Initialization**

Initialize the threshold values $\theta_j$, a positive constant learning rate α, with random number within the range [-2.4/$F_i$, 2.4/$F_i$], where $F_i$ is the maximum number of inputs connected to the single neuron.

**Activation**

Activate the input layer by presenting all samples repeatedly to the classifier. Set the desired output layer and then calculate the actual output of the output layer.

a) Calculate the actual output of the neuron in the hidden layer(s) at $p^{\text{th}}$ training cycle:

$$y_j(p) = R\left(\sum_{i=1}^{n} x_i(p) w_{ij}(p)\right)$$ ; R is the sigmoid function

b) Calculate the output of the neuron of the output layer:

$$y_k(p) = R\left(\sum_{j=1}^{m} y_j(p) w_{jk}(p)\right)$$

**Training**

a) Compute error signals at the output and the hidden layers (back propagation step):

$$e_k(p) = y_k(p) - d_k$$

$$\delta_k(p) = y_k(p)[1 - y_k(p)]e_k(p)$$

$$\delta_j(p) = y_j(p)[1 - y_j(p)]\sum_{k=1}^{l} \delta_k(p) w_{jk}(p)$$

b) Update the weights related to hidden and output layers:

$$w_{jk}(p+1) = w_{jk}(p) - \alpha\delta_k(p)y_j(p)$$

$$w_{ij}(p+1) = w_{ij}(p) - \alpha\delta_j(p)x_i(p)$$

**Iteration**

Increase iteration $p$: $p = p + 1$ process until convergence.

**Fig. 5.** Flowchart of Back Propagation Algorithm.

## 4 Experimental Evaluation

We develop an efficient image spam classification system based on image features using back propagation neural network. A histogram based method is used for visual features extraction. The file features of the image are selected based on the file type, file size and dimension of the image file. Experimental evaluations demonstrates the effectiveness of the propose system. To test our algorithm, we use a benchmark data set developed by G. Fumera *et al.* [17]. The corpora contains 5087 images combined of 3209 spam and 1878 non-spam images.

We evaluate our system by estimating three performance measures: Accuracy (A), Precision (P), and Recall (R). The measures can be defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

$$\Pr ecision = \frac{TP}{TP + FP} \tag{8}$$

$$\operatorname{Re} call = \frac{TP}{TP + FN} \tag{9}$$

where,

*TP* (true positive) = No of spam emails and identified as spam,
*FP* (false positive) = No of non-spam emails but identified as spam,
*TN* (true negative) = No of non-spam emails and identified as non-spam,
*FN* (false negative) = No of spam emails but identified as non-spam.

False positives are generally considered to be more harmful than false negatives. Therefore, our target is to ensure the low false alarm rate. If the value of precision is high, it obviously indicates that the false negative is high. In other words, the detector has misclassified many spam messages as legitimate (non-spam) message. On the other hand, a high recall indicates that the false positive is high, i.e. many legitimate messages are misevaluated as spam. We concern about the trade-off that exists between the spam and non-spam when we consider precision and recall values.

Table 2 illustrates the Signal to Noise ratio (SNR) for spam and non-spam image of GIF, JPEG and PNG format. Based on the SNR obtained for different features of an image it is possible to isolate spam message from the legitimate message. By analyzing our test dataset we find that most of the spam images in e-mails are GIF and non-spam images are JPEG type. A comparison of the performance between our proposed technique and other methods is reported in Table 3. Experimental results confirm that our proposed spam detection technique gives better performance comparable to existing methods.

**Table 2.** File features of an image

| File features | JPEG | GIF | PNG |
|---|---|---|---|
| $f_1$ | 0.268 | 0.192 | 0.498 |
| $f_2$ | 0.298 | 0.144 | 0.273 |
| $f_3$ | 0.010 | 0.032 | 0.312 |
| $f_4$ | 0.283 | 0.131 | 0.625 |
| $f_5$ | 0.312 | 0.803 | 0.451 |
| $f_6$ | 0.271 | 0.545 | 1.489 |

**Table 3.** Performance comparison of the proposed system with other techniques.

| Measures | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|
| Naïve Bayes | 94.53 | 83.15 | 96.65 |
| SVM | 95.09 | 96.38 | 97.04 |
| BPNN (proposed) | 97.89 | 93.75 | 98.02 |

# 5 Conclusion

In this paper, we present an efficient and robust method for image spam classification using back propagation neural network. The system analyzes the file features of the embedded image and extract the low level visual features as well. A gradient histogram based algorithm is utilized to extract the color feature points from the image. The extracted file features as well as the visual features are feed forwarded to the BPNN classifier to train the network. Experimental results confirms the effective performance of our proposed system comparable to the state-of-the-art methods. The results show the performance near to 98% accuracy and 0.03 false positive rate. Our future plan is to improve the algorithm to develop a complete classification system that is also capable of detecting textual spam image.

# References

1. Das, M., Prasad, V.: Analysis of an Image Spam in Email Based on Content Analysis. International Journal on Natural Language Computing (IJNLC) **3**(3), 129–140 (2014)
2. Patil, D., Turukmane, A.: Design and Development of Decision Making Model for Spam email Classification Using Neural Network. International Journal on Recent and Innovation Trends in Computing and Communication **3**(2), 327–330 (2015)
3. Islam, M.R., Zhou, W., Choudhury, M.U.: Dynamic feature selection for spam filtering using support vector machine. In: IEEE/ ACIS (ICIS) (2007)
4. Islam, R., Zhou, W.: An adaptive model for spam filtering using machine learning algorithms. In: 7th Int. Conference on Algorithms and Architecture for Parallel Processing (ICAAPP), Hangzhou, China (2007)
5. Sasaki, M., Shinnou, H.: Spam detection using text clustering. In: IEEE Proceedings of the International Conference on Cyber Worlds (2005)

6. Deshpande, V.P., Erbacher, R.F., Harris, C.: An evaluation of naïve bayesian anti-spam filtering techniques. In: Proceedings of the IEEE Workshop on Information Assurance, pp. 333–340 (2007)
7. Liu, Q., Qin, Z., Cheng, H., Wan, M.: Efficient modeling of spam images. In: 2010 Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI), pp. 663–666 (2010)
8. Mehta, B., Nangia, S., Gupta, M., Nejdl, W.: Detecting image spam using visual features and near duplicate detection. In: Proceedings of the 17th International Conference on World Wide Web, pp. 497–506. ACM (2008)
9. Li, P., Yan, H., Cui, G., Du, Y.: Integration of Local and Global Features for Image Spam Filtering. Journal of Computational Information Systems **8**(2), 779–789 (2012)
10. Wang, C., Zhang, F., Li, F., Liu, Q.: Image spam classification based on low-level image features. In: Proceedings of the ICCCAS, pp. 290–293 (2010)
11. Krasser, S., Tang, Y., Gould, J., Alperovitch, D., Judge, P.: Identifying image spam based on header and file properties using C4.5 decision trees and support vector machine. In: IEEE Workshop on Information Assurance (2007)
12. Liu, T., Tsao, W., Lee, C.: A high performance image-spam filtering system. In: Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science 2010, pp. 445–449 (2010)
13. Lai, C.-C., Wu, C.-H., Tsai, M.-C.: Feature selection using particle swarm optimization with application in spam filtering. Int. Journal of Innovative Computing **5**(2), 423–432 (2009)
14. Koprinska, I., Poon, J., Clark, J., Chan, J.: Learning to classify e-mail. Information Sciences **177**(10), 2167–2187 (2007)
15. Meyer, T.A., Whateley, B.: Spam bayes: effective open-source, Bayesian based, email classification system. In: First Conf. on Email and Anti-Spam (CEAS) (2004)
16. Drucker, H., Wu, D., Vapnik, V.N.: Support vector machines for spam categorization. IEEE Trans. on Neural Networks **10**(5), 1048–1054 (1999)
17. Fumera, G., Pillai, I., Roli, F.: Spam Filtering based on the Analysis of Text Information Embedded into Images. Journal of Machine Learning Research (Special Issue on Machine Learning in Computer Security) **7**, 2699–2720 (2006)
18. Wu, C.-T., Cheng, K.-T., Zhu, Q., Wu, Y.-L.: Using visual features for anti-spam filtering. In: Proceedings of the IEEE International Conference Image Processing, vol. 3, pp. 501–504 (2005)
19. Aradhye, H.B., Myers, G.K., Herson, J.A.: Image analysis for efficient categorization of image-based spam e-mail. In: 8th International Conference on Document Analysis and Recognition (ICDAR 2005), vol. 2, pp. 914–918 (2005)
20. Dredze, M., Gevaryahu, R., Elias-Bachrach, A.: Learning fast classifiers for image spam. In: Proceedings of the 4th Conf. Email Anti-spam (CEAS) (2007)
21. Wang, Z., Josephson, W., Lv, Q., Charikar, M., Li, K.: Filtering image spam with near-duplicate detection. In: Proceedings of the 4th Conf. Email Anti Spam (CEAS) (2007)
22. Biggio, B., Fumera, G., Pillai, I., Roli, F.: Image spam filtering by content obscuring detection. In: 4th Conference on Email and Anti-Spam (CEAS) (2007)
23. Zhang, C., Chen, W.-B., Chen, X., Tiwari, R., Yang, L., Warner, G.: A Multimodal Data Mining Framework for Revealing Common Sources of Spam Images. Journal of Multimedia **4**(5), 313–320 (2009)
24. Bowling, J.R., Hope, P., Liszka, K.J.: Spam image identification using an artificial neural network. In: MIT Spam Conference (2008)