ISSN: 0067-2904

Weighted k-Nearest Neighbour for Image Spam Classification

Ahmad M. Salih*, Ban N. Nadim
Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq

Received: 4/3/2020 Accepted: 16/6/2020

Abstract

E-mail is an efficient and reliable data exchange service. Spams are undesired e-mail messages which are randomly sent in bulk usually for commercial aims. Obfuscated image spamming is one of the new tricks to bypass text-based and Optical Character Recognition (OCR)-based spam filters. Image spam detection based on image visual features has the advantage of efficiency in terms of reducing the computational cost and improving the performance. In this paper, an image spam detection schema is presented. Suitable image processing techniques were used to capture the image features that can differentiate spam images from non-spam ones. Weighted k-nearest neighbor, which is a simple, yet powerful, machine learning algorithm, was used as a classifier. The results confirm the effectiveness of the proposed schema as it is evaluated over two datasets. The first dataset is a real and benchmark dataset while the other is a real-like, modern, and more challenging dataset collected from social media and many public available image spam datasets. The obtained accuracy was 99.36% and 91% on benchmark and the proposed dataset, respectively.

Keywords: E-mail, Image analysis, Image spamming, Weighted K-NN

.

الجار الاقرب الموزون لتصنيف صور البريد الالكتروني المزعجة

احمد مهدي* ، بان نديم
قسم الحاسبات، كلية العلوم، جامعة النهرين، بغداد، العراق

الخلاصة

تتسم خدمة البريد الإلكتروني لتبادل البيانات بالكفاءة والموثوقية. الرسائل العشوائية هي رسائل بريد إلكتروني غير مرغوب فيها يتم إرسالها بشكل عشوائي وبكميات كبيرة عادةً لأغراض دعائية. يعد تشويه الصورة المزعجة إحدى الحيل الجديدة لتجاوز مرشحات البريد العشوائي المستندة على تحليل النصوص او التعرف الضوئي على الحروف. يتميز اكتشاف الصور غير المرغوب فيها استنادًا إلى ميزات الصورة المرئية بالكفاءة من حيث تقليل تكلفة المعالجة وتحسين الأداء. في هذه المقالة، يتم تقديم طريقة للكشف عن صور البريد المزعج .تم استخدام تقنيات معالجة الصور المناسبة لالتقاط خصائص الصورة التي يمكن أن تميز الصور غير المرغوب فيها عن تلك الأصولية. تم استخدام خوارزمية الجار الأقرب الموزون كمصنف، وهو عبارة عن خوارزمية تعلم الة بسيطة ولكنها قوية. تؤكد النتائج فعالية الطريقة المقترحة حيث يتم تقييمها على قاعدتي بيانات . أول قاعدة بيانات عبارة عن مجموعة بيانات حقيقية ومعيارية بينما الأخرى عبارة عن مجموعة صور تشبه الحقيقة لكنها اكثر حداثة واكثر تحديًا تم جمعها عن طريق وسائل التواصل الاجتماعي

*Email: ahmadtoday2@gmail.com

والعديد من البيانات العامة المتاحة . كانت الدقة التي تم الحصول عليها هي 99.36٪ و 91٪على قاعدتي البيانات المعيارية و المقترحة على التوالي.

1. Introduction

E-mail is a reliable and popular communication medium that provides a free, or very cheap, and fast service. A Modern e-mail has many powerful capabilities, such as messaging with attachments, hyperlinks, and embedded images [1]. However, beside all the benefits of e-mail, the popularity and publicity of e-mails makes it an attractive goal to misuse. Spamming is an example of such misuse. Spamming is the utilization of e-mail systems to send unsolicited messages called spams, especially for advertising [2]. According to Symantec recent statistics, spam is accounted for approximately 50% of all e-mail traffic [3]. As a reaction to spam, several anti-spam filters have been proposed. The first generation of spam was in textual form. Keyword-based spam filters are efficient in detecting textual spam. Spammers, however, developed new tricks, such as image spam where the spam text is embedded within an image. As a simple solution for image spam, OCR was used to convert an image's textual content into plaintext format, and then keyword-based filters can be used to identify spam from non-spam (or ham) texts. To make OCR useless, spammers use obfuscation tricks (such as adding noise, complex background, etc.) with a goal of making the spam image readable by humans but unreadable by machine. This has led to a new generation of spam filters based on image visual characteristics [4]. Figure -1 shows examples of spam images.

Figure 1- Examples of spam images (ISH dataset).

2. Literature review

Several research studies have been published in the field of interest. The following selected studies are the most interesting and recent ones.

Kumaresan *et al.* [5] proposed a filter for detecting image spam based on color features and image file properties. Specifically, the authors depend on the RGB (Red-Green-Blue) and HSV (Hue-Saturation-Value) histograms as features. K-nearest neighbor (k-NN) was used as a classifier. In that research, k-NN yielded an accuracy of 94.5% on spam archive dataset.

Annadatha *et al.* [6] introduced a spam detection approach using Support Vector Machine (SVM) . The proposed approach utilizes the following features: first order moments of RGB histograms, local binary pattern (LBP), histogram of oriented gradients (HOG), and total number of edges, in addition to image file features. These features accounted for 21 features. Recursive Feature Elimination using SVM weights was used as a feature reduction technique. By using all the 21 features, the linear SVM achieved an accuracy of 96% on image spam hunter (ISH) dataset, while by using only 13 features; it achieved an accuracy of 97.25% on the same dataset. It is obvious that eliminating the number of features improves the performance and reduces the time cost.

Chavda *et al.* [7] proposed an image spam detection system using SVM classifier and based on a wide variety of features , include all features presented in [6] in addition to entropy and first order moments of HSI color histogram. The total number of used features was 38 features. They achieved accuracy values of 97% and 98% on ISH and dredze datasets, respectively. The results showed that the usage of more features does not improve the performance significantly, because some image features do not capture the special characteristics of spam images.

Kumar *et al.* [8] proposed a convolution neural network (CNN)-based image spam classification approach. Their CNN consists of three convolutional layers connected to max-pooling window for dimensionality reduction, while dropout was used for regularization. The proposed approach was trained and tested against ISH dataset and achieved an accuracy of 91.7%. The relatively low performance obtained by using CNN was due to the small size of the dataset and ignoring the image file features.

Singh *et al.* [9] introduced neural network and deep neural network network-based spam image classifiers. They used 38 features that included Meta data, color (RGB and HSV), texture, shape, and noise features. The best achieved accuracy was 99.07% using neural network, with 10 fold cross-validation on ISH dataset. While, for deep neural network the best accuracy was 98.78. The experimental results showed that using neural network with adequate number of discriminative features outperforms all other proposed approaches. Other researchers used the same features on the same dataset (ISH) with SVM and yielded an accuracy of 97% [7].

Yang *et al.*[10] proposed a multi-modal spam detection architecture based on model fusion to detect whether the spam is hidden in the text or in the image. They employed the Long Short-Term Memory (LSTM) and CNN models, individually, to analyze both text and image parts of the e-mail and obtain two classification probability values. Then, the two values were fused to determine whether the e-mail is spam or non-spam. The proposed model's performance was measured against a hybrid dataset that consisted of text spams (Enron dataset) and image spams (spam archive and personal spam/ham dataset). The average accuracy values were 98%, 92%, 98% on text, image, and hybrid datasets, respectively. Their proposed model achieved high accuracy values on text and hybrid (image/text) spam, with relatively low performance when only spam images were tested.

Sharmin *et al.* [11] presented image spam classifiers using SVM, multilayer perceptron (MLP) and CNN. For SVM and MLP, canny edge detector was used to extract efficient edge information. The CNN classifier consisted of three convolutional layers, three max-pooling layers, and a dropout unit to avoid overfitting. Their accuracy results were 98.7, 95.5, and 99.02 using SVM, MLP and CNN, respectively, on ISH dataset.

3. Methodology

This section presents the theoretical background of image processing and machine learning techniques used throughout this paper.

3.1 Image processing

Spam images have some special characteristics in terms of color distribution and the amount of texture information. Image features are analogous to image characteristics. Several image processing techniques can be used to extract distinctive features that can differentiate spam from non-spam images. Some of the most useful techniques are:

Dominant color descriptor (DCD): an image is visually understandable depending on a few main colors, while other colors are either for details or noise so they are not inherent and can be neglected [12]**.**

Gray level co-occurrence matrix (GLCM): The GLCM describes the image texture depending on the number of image pixel pairs with certain intensity values arranged in certain spatial relationships [13] .

Local binary pattern (LBP): LBP is a simple, yet effective, image texture descriptor which labels the image's pixels by thresholding the neighboring pixels depending on the value of the current pixel. LBP descriptors efficiently capture the local spatial structures and the contrast in gray level images [14].

Color moments: Descriptive statistical measures are a useful data analysis tool which could provide an accurate summary of the data. An image is a set of pixel values. Therefore, it can be characterized by the mean, variance, skewness, kurtosis ,etc., which are called color moments [12].

Hue , Saturation , Value (**HSV) color space:** HSV color space takes some advantages over RGB color, one of which is decoupling the chrominance/ luminance components from each other [15] . This makes it desirable for image analysis, especially for the purposes of image spam detection where the variation in brightness (luminance) in non-spam images is very high as compared to that of spam images [6, 16].

3.2 High level features

According to image spam literature, image features which are extracted using image processing techniques are called low-level features. Whereas high-level features are general properties of an image file, such as its size, width , height, aspect ratio, file format, compression ratio, filename, bit depth, and signal to noise ratio (SNR) [17]. The following equations can be used to compute image properties that include more than one variable:

*Compression ratio = (imge width*image height*bit depth) / (image size)* *Eq.1 reference [6]*

Aspect ratio = (image width) / (image height) *Eq.2*

SNR = (mean pixel value (μ)) / (standard deviation of the image pixel values (σ)) *Eq.3*

3.3 Weighted K-Nearest Neighbour (Weighted K-NN)

K-NN is a non-parametric, instance-based, and lazy classifier. The laziness of K-NN is due to the lack of learning stage. Instead, K-NN stores all available training data and classifies the new test instance based on a distance metric [18]. KNN uses a simple majority voting method for predicting the class of the test instance. It is very susceptible to unbalanced data. To improve this, an improved approach is to weight the votes of k nearest neighbors according to their distance from test instance. In Weighted K-NN, the closer neighbours obtain heavier weights than the farther ones [19]. Figure - 2 presents examples of K-NN and Weighted K-NN. The predicted class label of test instance x is a triangle if K-NN is used, while it is a rectangle if Weighted K-NN is used. The weight of each neighbour can be computed using the following equation:

Weight =1/distance *Eq.4*

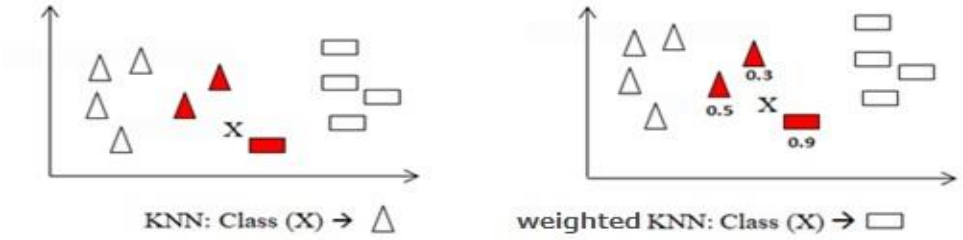


Figure 2- Weighted K-NN examples [19]

The seignificant advantage of K-NN that makes it suitable for nonlinear real data classification is its rich hypothesis space, i.e. the set of local functions that the K-NN is able to select as being the solution [20].

3.4 One- Ham Neighbor (1-HN)

1-HN is a slightly modified version of K-NN. Instead of using the majority voting or calculating the weights of the k neighbors, 1-HN assigns the new image to the ham class if at least one of the k nearest neighbors is ham. The goal is to reduce the false positive rates.

3.5 Evaluation metrics

In the context of spam detection, the proposed techniques were evaluated based on accuracy, ROC curve, and False Positive Rate. The term True Positive (TP) gives the number of correctly classified spam e-mails, while True Negative (TN) is the number of non-spam e-mails that are correctly classified. The term False Positive (FP) represents the number of non-spam e-mails identified as spam, while False Negative (FN) is the number of spam e-mails that are miss-classified as non-spam [2]. Accuracy is given in terms of TP, FP, TN and FN as

Accuracy = TP + TN / (TP +TN+ FP + FN) *Eq.5*

For any binary classifier, ROC curve is the plotting of True Positive Rate (TPR) against False Positive Rate (FPR) for various threshold values. TPR is also called sensitivity, while TNR is called specificity. FPR value is calculated via the relation FPR = 1 - specificity. Area under the Curve (AUC) value of the ROC curve determines the efficiency of the classifier. An AUC value of 1 denotes ideal classification with zero FP and FN.

TPR and TNR can be computed as follows:

$$TPR = TP / TP + FN \text{ and } TNR = TN / TN + FP \quad Eq.6$$

False positive rate is an important criterion to evaluate spam filter, as it is acceptable to mark a spam image as non-spam but it is not acceptable to mark a non-spam image as spam, because this leads to the loss of legitimate information [17].

4. Proposed image spam detection schema

This section provides the implementation details of the image processing techniques mentioned in the previous sections.

4.1 Feature extraction

i- Color features

Color moments (Variance, skewness , kurtosis) : a 100- probabitity histogram is built for each channel of RGB and HSV channels , then the three color moment features are calculated from each histogram , resulting in 18 color features.

Dominant color discriptor (DCD): to obtain dominant colors of an image, a 256-graylevel histogram is built, then only bins with values above a certain threshold are considered as dominant colors.

ii- Texture features

Gray level co-occurrence matrix (GLCM): entropy and homogeneity features are extracted from GLCM with θ=0,90 and unit distance, and used as two texture features.

Local binary pattern (LBP): a histogram of LBP image is built , then the entropy of LPB histogram is considered as the third texture feature**.**

Figures- 3 and 4 show the differneces between spam and non-spam images in terms of the amount of texture information and color distributions.

Figure 3- LBP opertaor results on spam image (left) and non-spam image (right).

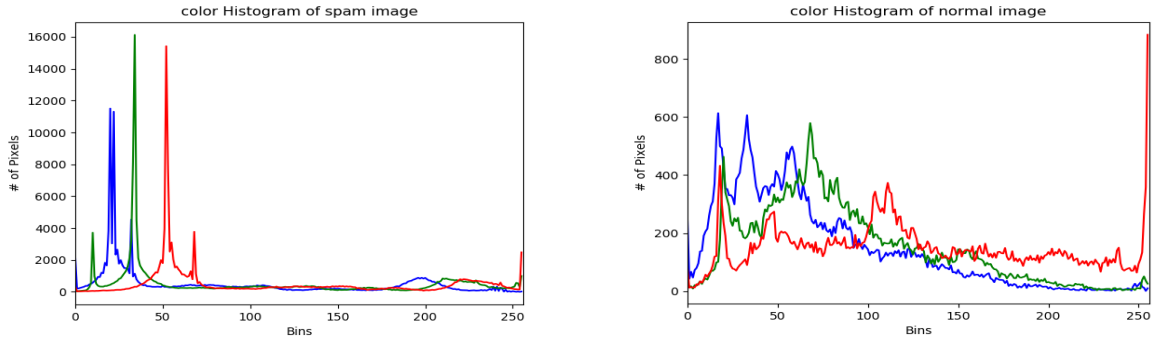


Figure 4- Spam image RGB histogram (left) and non-spam image RGB histogram (right).

iii- High level features

Image file properties, which include compression ratio, aspect ratio, SNR , image area , image size and image dimensions, were used as high level features. The total number of the selected features was 29 . The extracted features were scaled using z-score normalization and then fed to the classifier .

5. Datasets

Two datasets were considered in this paper. One of these datasets is a public dataset that consists of real spam and non-spam e-mails, while the other is collected from various sources. All images in both datasets are in JPEG format.

i. Image Spam Hunter (ISH)

This dataset was collected by authors of a previous paper entitled "Image Spam Hunter" [21] . It consists of 929 real spam and 810 non-spam images.

ii. The proposed dataset

The spam images are selected from social media advertisemnets. To ensure the diversity of spamming tricks , a selected set of spam images from three public spam image datasets (spam archive , princiton benchmark, and dredze) was included. The selection criterion is to include spam images whose visual features are similar to the features of non-spam images. The idea behind creating such dataset is to build a challenge dataset that can fool the existing image spam classifiers. The proposed dataset is composed only of spam images (n = 892).For the experiments, ISH non-spam images were used with this dataset. This proposed image spam dataset is available for download at the the link provided in a previous study [22] .

6. The Results

This section reports the experimental results obtained in the present work. All the experiments were conducted on a Windows 10 Machine with 4 GB RAM and ci5 processor . This project was implemented using Python programming language .

6.1 Feature analysis

For each image , 29 informative features were extracted .The used features reflect the discriminative visual statistics of spam image as compared to the non-spam image. Figures-(5 and 6 present samples of the distributions of 4 features for each dataset.

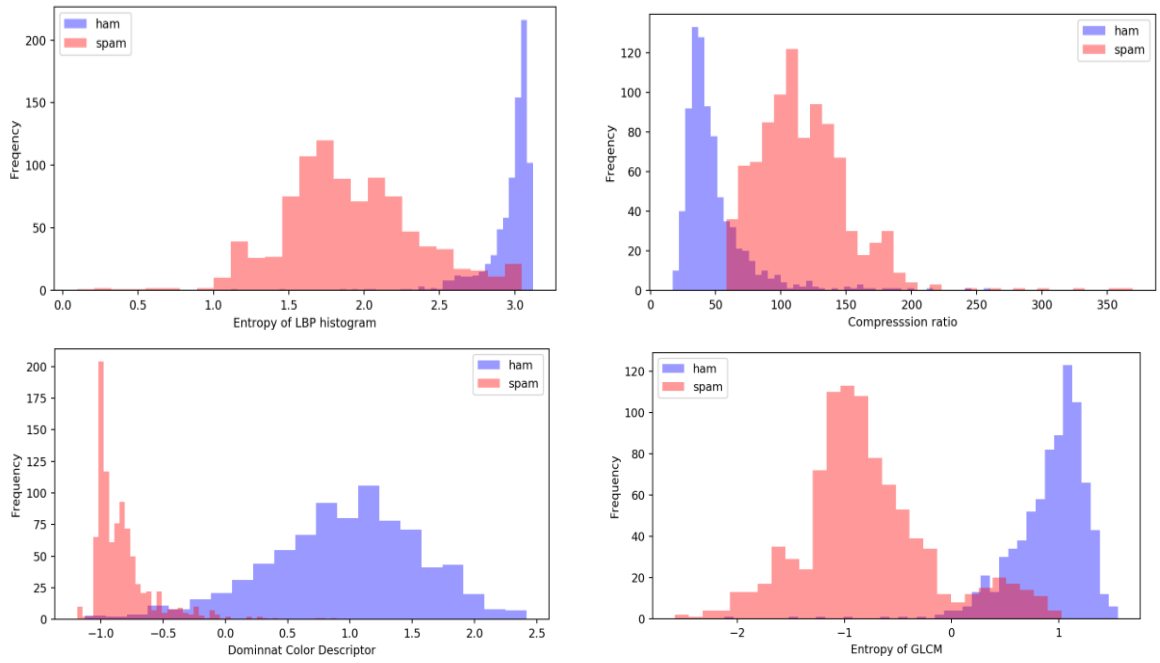


Figure 5- Feature distribution for spam against ham images on ISH dataset

For each of the features in Figure -5, there is an acceptable separation between the spam and non-spam distributions. Therefore, it is expected that the results based on these features would distinguish between spam and non-spam images with acceptable accuracy.

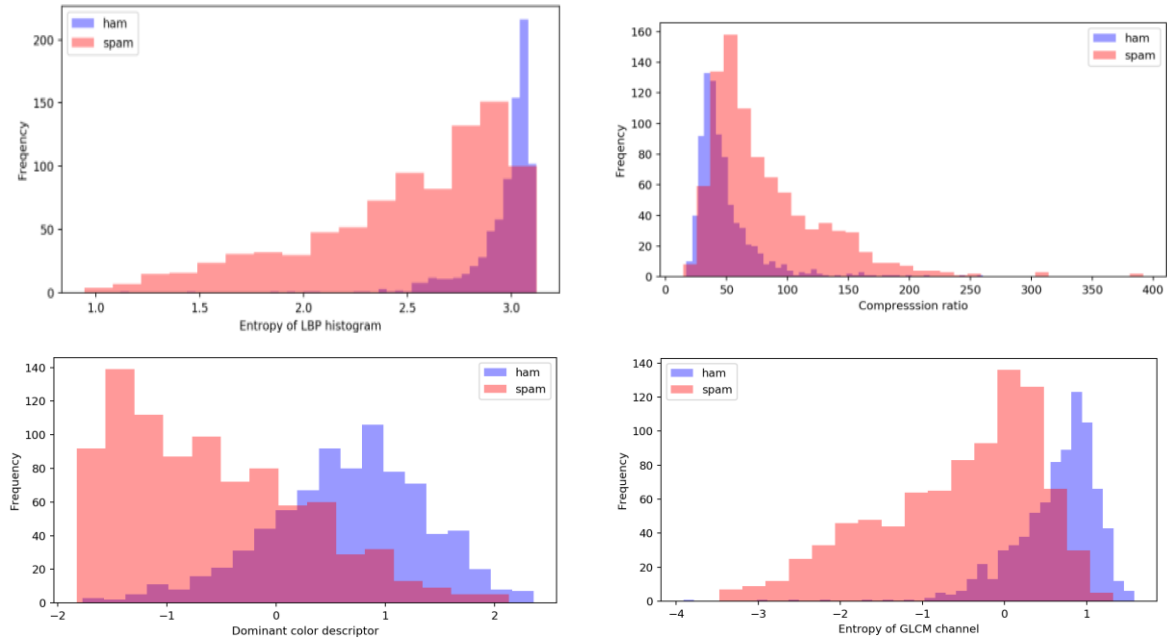


Figure 6- Feature distribution for spam against ham images on the proposed dataset

From figure -6, it is noticeable that the separation capabilties of the presented features were reduced. Therefore , the purpose of creating a challenge image spam dataset is satisfied.

6.2 Classification

K-NN has two parameters ; the fisrt one is the K value whereas the other is distance metric. To evaluate the proposed visual-based image spam filter, Weighted K-NN and 1-HN with different k values and distance metrics are considered for the experiments. A ten-fold cross-validation is adopted to produce ten distinct folds. For the ten-iterations, one fold is used as a testing set and the other nine folds are used as a training set. The average accuracy of testing sets is the accuracy of the classifier.

i. Weighted K-NN

Table -1 provides the obtained results of weighted K-NN with K=3 for each distance metric over the two datasets under consideration.

Table 1- Average accuracy and false positive rate for each metric

Dataset	Distance metric	Average Accuracy	Average FPR
ISH	Manhattan	99.13%	0.014
	Chebyshev	99.19%	0.013
	Euclidean	99.36%	0.011
Proposed	Manhattan	91%	0.08
	Chebyshev	88.6%	0.13
	Euclidean	90.2%	0.10

Based on Table -1, the proposed method is able to achieve a prominsing performace with accuracy values of 99.3% on ISH dataset and 91% on the proposed dataset.

Figures-(7 and 8) present the results of the proposed classifier in the form of ROC curves. The corresponding mean AUC values are 0.99 and 0.92 on ISH dataset and the proposed dataset, resepectively.

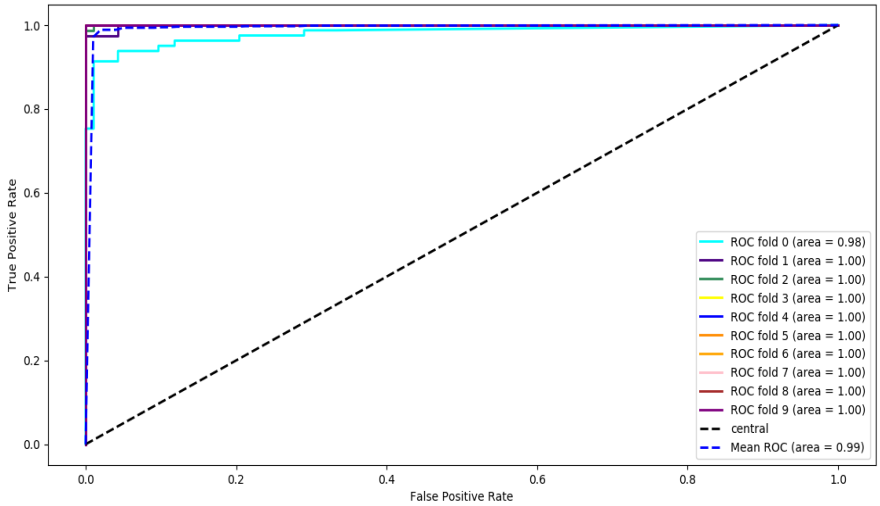


Figure 7- ROC curves on ISH dataset

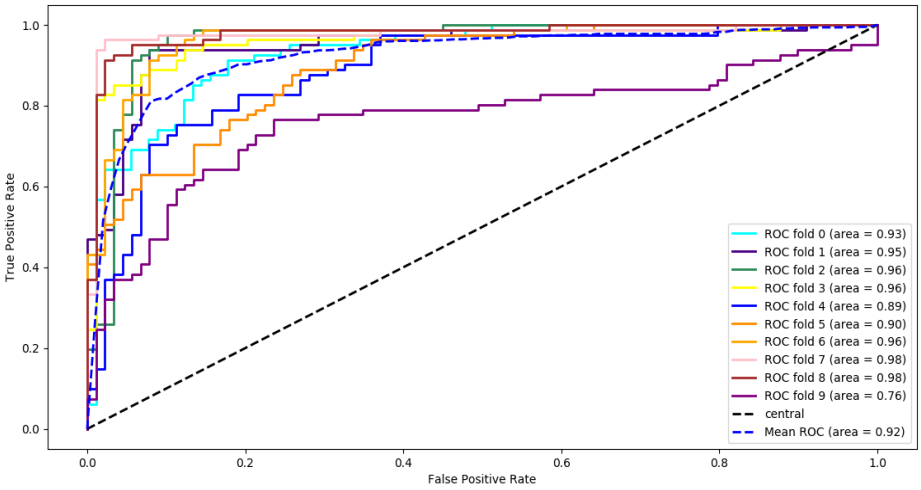


Figure 8- ROC curves on the proposed dataset

Figure -9 shows that using k= 5 gave the best accuracy among he other possibilities.

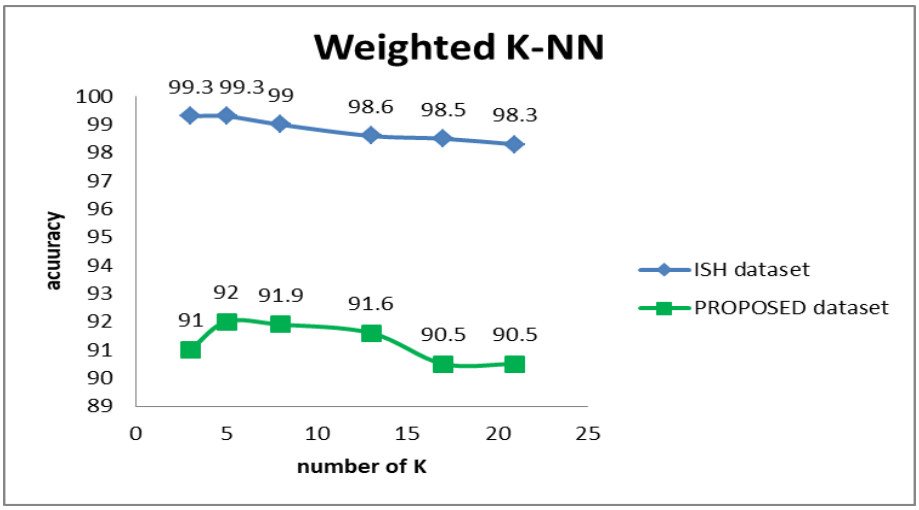


Figure 9- Weighted K-NN results on ISH and proposed datasets with different value of k

ii. 1-HN

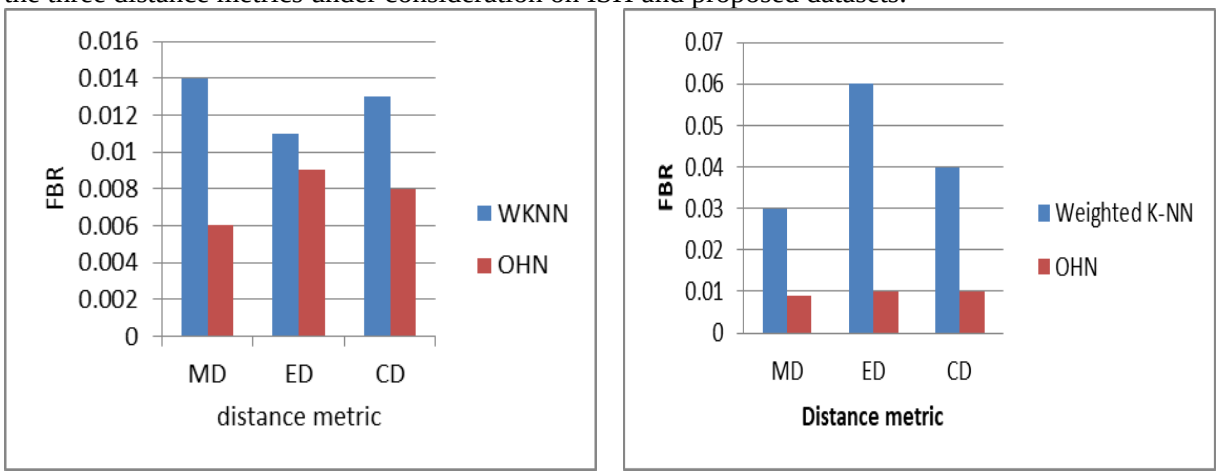
Table -2 provides the obtained results using 1-HN with k=3 for each distance metric over the two datasets under consideration.

Table 2 – Average accuracy and false positive rate for each distance metric

Dataset	Distance metric	Average Accuracy	Average FPR
ISH	Manhattan	98.8%	0.006
	Chessboard	98.6%	0.008
	Euclidean	98.9%	0.009
Proposed	Manhattan	83.1%	0.02
	Chessboard	82.9%	0.05
	Euclidean	83.4%	0.04

As the goal of 1-HN is to reduce the FPR value , table - 2 shows that 1-HN was able to achieve its goal with FPR values of 0.6% and 2% on ISH and the proposed dataset, respectively. However, 1-HN achieves this goal at the expense of the overall accuarcy .

Figure -10 provides a FPR comparison of the two classifiers (i.e., Weighted k-nn versus 1-HN) over the three distance metrics under consideration on ISH and proposed datasets.



A B

Figure 10- Comparison of FPR for Weighted K-NN and 1-HN on ISH (A) and proposed dataset (B)

7. Conclusions

Spamming can undermine the efficiency of e-mails. Spam e-mails can be also used as a bridgehead for other serious kinds of cyber-crimes, such as phishing. Although there are several attempts to stop spam, this task is always a challenge as there is no clear criterion to distinguish legitimate e-mails from spam ones.

In this paper, an efficient and robust image spam method is presented. The method analyses the file properties of the image and extracts the low-level visual features. Weighted K-NN as a machine learning classifier was implemented using different distance metrics and different k values. Due to lack of modern public datasets for image spam research, a tricky image spam dataset is proposed. Experiments based on two datasets (public and proposed) demonstrated the effectiveness of the proposed method. The obtained results for public dataset outperform those obtained in the previous works. The primary challenge facing the present work is the high computational cost of K-NN during classification. Future works are to overcome the high computational cost of K-NN and to develop a complete spam classification system that is capable of detecting spams whether in image-form or text-form.

References

1. James F. and Kurose, K.W.R. **2016** *Computer networking : a top-down approach*.7th edition, publisher :Pearson Education.
2. Attar, A., R.M. Rad, and R.E. Atani. **2013**. *A survey of image spamming and filtering techniques. Artificial Intelligence Review,*. **40**(1): 71-105. link : https://doi.org/10.1007/s10462-011-9280-4.
3. Dada, E.G., et al. . **2019**. *Machine learning for email spam filtering: review, approaches and open research problems.* Heliyon,. **5**(6). link : https://doi.org/10.1016/j.heliyon.2019.e01802
4. Dhavale, S.V. . **2017**. *Advanced image-based spam detection and filtering techniques*. Information Science Reference. DOI: 10.4018/978-1-68318-013-5
5. Kumaresan, T., S. Sanjushree, and C. Palanisamy. **2014**. Image spam detection using color features and K-Nearest neighbor classification. *Int. J. Comput. Inf. Syst. Control Eng.* **8**(10): 1746-1749.
6. Annadatha, A. and M. Stamp . **2016**. Image spam analysis and detection. *Journal of Computer Virology and Hacking Techniques*. **14**(1): 39-52.link: https://doi.org/10.1007/s11416-016-0287-x
7. Chavda, A. . **2017**. Image Spam Detection, Master thesis in *computer science* , San Jose State University. link : https://doi.org/10.31979/etd.myqt-f92r
8. Dinesh Kumar, A. and S. KP, **2018**, DeepImageSpam: Deep Learning based Image Spam Detection. arXiv preprint arXiv:1810.03977,.
9. Singh, A.P. . **2018**. Image Spam Classification using Deep Learning, Master Thesis in *computer scinece*. San Jose State University.link : https://doi.org/10.31979/etd.wehw-dq4h
10. Yang, H., et al. . **2019**. A spam filtering method based on multi-modal fusion. *Applied Sciences*. **9**(6): 1152. DOI: 10.3390/app9061152
11. Sharmin, T., et al. .**2020**.Convolutional neural networks for image spam detection. I*nformation Security Journal*: A Global Perspective p. 1-15. link : https://doi.org/ 10.1080/19393555 .2020. 17 22867
12. Zhang, D. **2019**. *Color Feature Extraction*, in *Fundamentals of Image Data Mining*. Springer. p. pp 49-80.
13. Hung, C.-C., E. Song, and Y. Lan . **2019**. *Image texture analysis*. Springer.
14. Matti Pietikäinen , A.H., Guoying Zhao , Timo Ahonen . **2011**. *Computer Vision Using Local Binary Patterns*. Vol. 40.: springer.
15. Alya'a, R.A. and B.N. Dhannoon . **2019**. Real Time Multi Face Blurring on Uncontrolled Environment based on Color Space algorithm. *Iraqi Journal of Science*. : 618-1626.
16. Gonzalez, R.C. . **2018**. *Digital Image Processing*. 4th edition.: Pearson.
17. Kumar, J., S. Taterh, and D. Kamnthania . **2018**. *Study and Comparative Analysis of Various Image Spamming Techniques*, in *Soft Computing: Theories and Applications*. Springer. p. 351-365.
18. Naoum, R.S. and Z.N. Al-Sultani . **2012**. Learning vector quantization (LVQ) and k-nearest neighbor for intrusion classification. *World of Computer Science and Information Technology Journal* (WCSIT). **2**(3): 105-109.
19. Dudani, S.A. . **1976**. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics* .(4): 325-327.
20. Goodfellow, I., Y. Bengio, and A. Courville, **2016.** *Deep learning*. MIT press.
21. Gao, Y., **2008** . *Image spam hunter*, in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE: Las Vegas, NV, USA.
22. *proposed image spam dataset*. [accessed: June 1, 2020]; Available from: https://www.dropbox.com/s/rgzqty186afwna8/the%20proposed%20spam%20image%20dataset.rar?dl=0.