ENSEMBLE OF MULTI-SPATIAL RESOLUTION FOR IMAGE SPAM FILTERING

Nor Azman Mat Ariff[1], Azizi Abdullah[2] and Mohammad Faidzul Nasrudin[2]

[1]Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

[2]Faculty of Technology and Information Science,
Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia.

Corresponding Author's E-mail: [1]nazman@utem.edu.my

Article History: Received 28 February; Revised 20 April 2023; Accepted 24 May 2023

ABSTRACT: Image spam is a type of spam e-mail that contains an image in the body of the e-mail and holds malware and other malicious threats. The rise of image spam has become a serious concern for e-mail users. This paper presents a spam image classification scheme with two primary goals. Firstly, Multi Spatial Resolution (MSR) with four different levels of resolution is proposed to improve the representation of images by incorporating spatial information between features. Due to the fact that MSR generates distinct image representations for each level, the predictions obtained from each representation may give different results. Thus, the final prediction of whether an image is spam or legitimate is difficult to determine. To solve this problem, an ensemble of MSR is proposed to combine the class probabilities of the model at each level to obtain a final prediction. The experiment was carried out on two public data sets, namely Dredze and SpamArchive. The results show that the classification accuracy improves as the level of MSR increases, outperforming the accuracy of level 0 that relies on global features alone. Meanwhile, the ensemble of MSR improved the accuracy of MSR and outperformed all four MSR models for both datasets.

KEYWORDS: *Multi Spatial Resolution; Ensemble Methods; Base64; N-Gram; SVM*

1.0 INTRODUCTION

Despite being one of the earliest Internet services, e-mail remains the most widely used communication method today, offering an efficient way to convey messages both formally and informally. However, e-mail services have also become a favored method for marketers to promote their products in bulk, taking advantage of the free bandwidth and storage available [1]. The situation is further worsened by the presence of malicious code, such as malware and viruses, which can be embedded in e-mails [2].

Spam images can be analyzed in the middle of the network using firewalls and at the endpoint detection system, i.e., at the e-mail server or user device. However, the machine learning community has primarily directed their research efforts towards endpoint detection systems using various techniques such as Optical Character Recognition (OCR) [3][4], content-based filtering using low-level features [1, 5-7], and the analysis of image metadata such as image size, file format [8], and Base64 codes [9]. However, spammers have started using obscuring techniques, which makes the OCR technique ineffective. While content-based filtering is effective in detecting image spam, it can be expensive as it uses low-level features that require extensive computational resources. Another problem is that most of these techniques rely on basic bag-of-features. This may limit their ability to describe features as they do not include spatial information. Spatial information can improve classification tasks by providing more precise feature representation. Spatial Pyramid Matching (SPM) is a method that incorporates global and local spatial information into a feature. Meanwhile, an ensemble method is a post-classification technique and is excellent in improving classification accuracy. It is very straightforward to implement and does not require high processing power.

This study's contribution can be summarized as follows: (a) Propose a Multi Spatial Resolution (MSR) technique using Base64 code as features. Base64 code was chosen as a feature to represent the image because it can be extracted directly from the e-mail when in the middle of the network such, as on a firewall or on an endpoint system such, as an e-mail server or a user device. The concept of SPM has been applied to images. In this study, inspired by SPM, MSR implemented a spatial pyramid to text. The reason for this is that images are encoded in the Base64 format, which results in their conversion into a textual representation. Similar to SPM, MSR can analyze images at different levels using different Base64 representations. By incorporating spatial information between features, MSR is expected to achieve superior classification performance at the multi-resolution level, in contrast to the basic level, which relies on global features. (b) Propose an ensemble of MSR using three combination methods, namely mean, weighted mean, and product. Due to the fact that MSR generates distinct image representations for each level, the predictions obtained from each representation may give different predictions. Thus, the final prediction of whether an image is spam or legitimate is difficult to determine. An ensemble of MSR is used to obtain higher accuracy in the final prediction compared to any model at the level of MSR.

2.0 RELATED WORK

2.1 Image Partitioning Scheme

The field of image classification presents a challenge to the image-processing community, and significant research has been conducted to identify features that can better represent images and achieve high detection rates. Each feature may provide a distinct representation. While a basic frequency of features can offer satisfactory levels of detection, extracting more information from these features can further enhance the detection rate. Different partitioning methods compute different histograms, leading to different image representations. Popular and widely used partitioning schemes include the global approach, local approach, and spatial pyramid approach.

A global feature provides a condensed representation of image content, resulting in a compact feature set [9, 10]. This approach does not implement image segmentation; thus, features are directly computed from the image. Due to the lack of information about the spatial arrangement of visual elements, the global approach was unsuccessful in accurately representing an image, resulting in poor prediction decisions.

Local features in images refer to the distinctive patterns or structures that exist within certain regions of an image [11, 12]. These features are computed based on the characteristics of small regions in an image, often referred to as image patches. The regions or patches can be extracted from different scales, orientations, and positions in the image, resulting in a set of local features with different characteristics.

In 2006, [14] first introduced the concept of spatial pyramid matching, which has been shown to produce notable enhancements in classification accuracy when compared to conventional bag-of-features methods. Referring to figure 1; this approach involves dividing an image into smaller sub-regions and computing histograms of visual features within each of these sub-regions. These histograms are then concatenated to generate a feature vector that represents the entire image. The motivation behind this approach is that images can be analyzed at different levels of detail, and combining features from different levels can result in a more informative representation of the image.

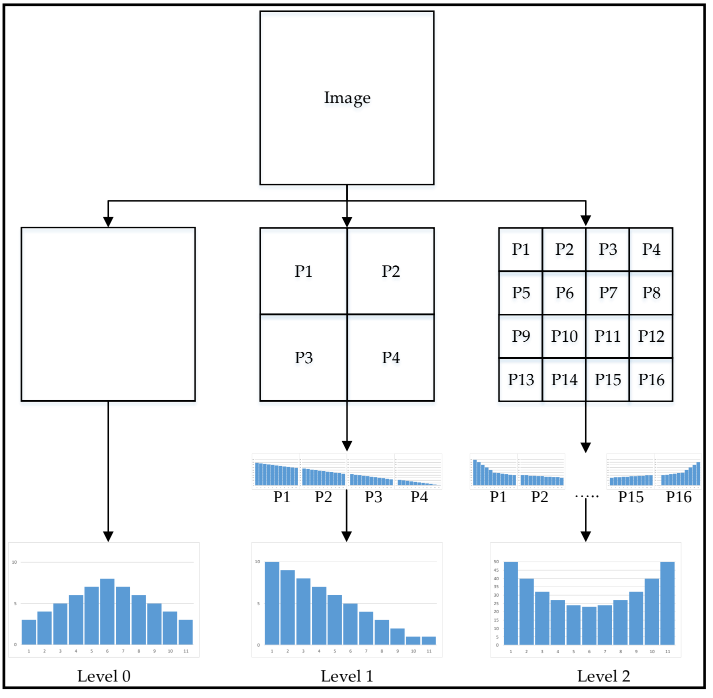

Figure 1 : The process of Spatial Pyramid Matching

Most researchers suggest that the highest level that the spatial pyramid should be processed is often limited to level 2 or level 3 [14, 15], as it typically provides the best recognition performance. The performance usually starts to decline at higher levels. The main advantage of the spatial pyramid approach is that certain images can be better represented using a combination of different levels or at a certain level rather than a global approach. Previous research has shown that combining feature vectors from multiple levels can lead to further improvements in classification performance compared to using a single level [17][16-18].

2.2 Ensemble Method

The ensemble method has been widely used since the 1990s. The goal of the ensemble method is to improve the generalization ability using multiple classifiers [19]. The ensemble method involves two main steps: first, creating classifiers from multiple learning algorithms, and second, combining these classifiers. To create a reliable ensemble method, it's important to ensure that the classifiers are both accurate and diverse [20].

Refer to figure 2, bagging and boosting are two types of widely used ensemble methods [20, 21]. They differ in how they generate and combine the weak model. Bagging, also known as the parallel ensemble method, is a widely used and early ensemble method where classifiers are built in parallel [22].

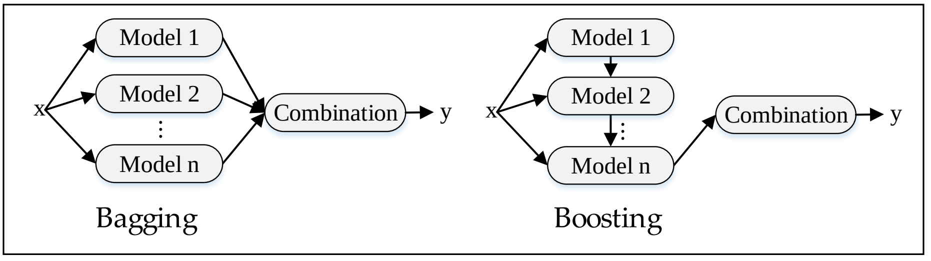

Figure 2 : Bagging vs Boosting

The boosting technique is a group of algorithms that can transform weak classifiers into strong ones by combining them in a sequential manner [21]. Boosting's main concept is to rectify previously misclassified instances by assigning them higher weights in the training process of the next classifier.

The combination method is crucial in enabling the ensemble method to attain effective generalization ability. Diverse combination methods have been used by researchers, which encompass mean rule, product rule, weighted mean rule, etc. Below is the formula to calculate the ensemble output for product rule(1), mean rule(2), and weighted mean rule(3).

$$H^j(x) = \prod_{i=1}^{T} h_i^j(x) \tag{1}$$

$$H^j(x) = \frac{1}{T} h_i^j(x) \tag{2}$$

$$H^j(x) = \frac{1}{T} w_i h_i^j(x) \tag{3}$$

Where: $H^j(x)$ = Ensemble output for class c_j
$h_i^j(x)$ = probability output for class c_j
T = number of individual classifiers
w_i = weight for $h_i^j(x)$ classifier.

3.0 ENSEMBLE OF MULTI-SPATIAL RESOLUTION

In this section, the proposed method, an ensemble of MSR will be described. It involves two main phases, as shown in figure 3, namely Multi Spatial Resolution (MSR) and the ensemble method.

In the first phase, spam and legitimate images is converted to Base64 code. MSR method decomposes the Base64 code into a three-level pyramid at different resolutions, namely levels 1, 2, and 3, in which each level contains different numbers of partitions. Based on figure 3, level 0 is the basic level where it is not partitioned and has only one partition labelled as P0. Global features are used to represent the image at this level and describe the entire image as a whole. At level 1, the base64 code is divided into two partitions, P1 and P2. Features on P1 describe specific partitions of P1 and are extracted locally.

Similarly, for P2 to P14, the features are extracted locally on their own partition. Since the Base64 code has converted the image to text problem, n-gram extraction can be used to represent the image using the n-gram of Base64 codes. The process of converting Base64 to n-gram uses these two steps. First, split the Base64 into individual words or characters, depending on the desired level of granularity (for example, n = 1..5). Then, generate all possible contiguous sequences of n items from the tokenized Base64. For example, if we want to extract trigrams (n = 3) from the Base64 code "R0lGODlhp", we would generate the following trigrams: "R01", "01G", "1GO", "GOD", "OD1", "D1h", and "1hp". These generated n-grams will be used as feature descriptors to describe the image. Feature descriptors must first be identified before the feature vector is generated. At level 0, the process of identifying feature descriptors involves performing multiple classification accuracy evaluations with various sizes of feature descriptors. The size of the feature descriptor that yields the highest accuracy is utilized to extract features from specific partitions ranging from P1 to P14. This means that all partitions from P0 to P14 use the same feature descriptor as at level 0. The feature vector for each partition can then be generated by computing the frequency of each n-gram based on the feature descriptor.

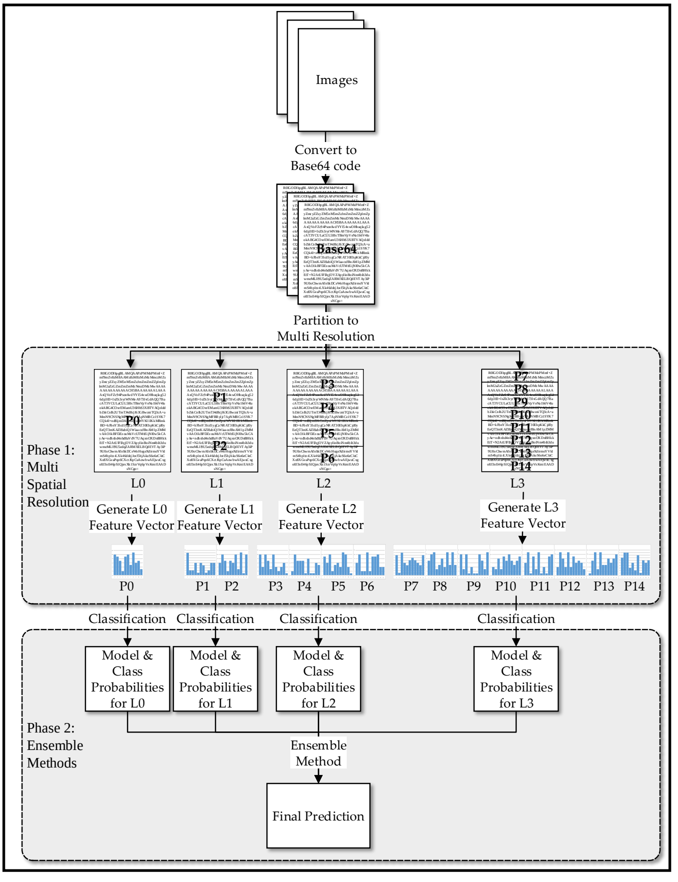

Figure 3: Block diagram of ensemble of Multi Spatial Resolution

Figure 3 shows that the generated histogram is different for each partition. This is because global feature histograms (P0) represent the overall distribution of n-gram features across the entire Base64 code, while local feature histograms (P1 to P14) represent the distribution of n-gram features within a specific partition in the Base64 code. The histogram generated from each partition is a feature vector that was then go through the classification process. The effectiveness of MSR in improving the classification accuracy of level 0 is evaluated by assessing the classification accuracy of each level.

In the second phase, ensemble methods are used to improve the classification accuracy of MSR. The effectiveness of an ensemble method depends on the diversity of the L0, L1, L2, and L3 models and the combination method used to aggregate their class probabilities. Classifiers should be fine-tuned to produce class probabilities during the classification process. During phase 1, these L0, L1, L2, and L3 models produce their own class probabilities as output and then is used as input to the ensemble method. The ensemble method combines the class probabilities of all the models to generate the final classification accuracy. This final accuracy is expected to be more accurate and robust than the accuracy of any single classification model in MSR. In this study, bagging with three combination methods is proposed, namely mean, weighted mean, and product. For the weighted mean, Particle Swarm Optimization (PSO) used to obtain the optimal weights for all four models. PSO is implemented on the training data, and the best weights obtained from the training data are used to predict the test data.

4.0 RESULTS AND DISCUSSION

In this section, we introduce the datasets used in our experiments, explain the setup and finally, report the results of the two datasets.

4.1 Dataset

Publicly available image spam datasets used in this study are Dredze and SpamArchive datasets. The Dredze dataset has both; spam and legitimate images. Since the SpamArchive dataset has spam images only, they will be combined with legitimate images from the Dredze dataset.

4.2 Experimental Setup

Repeated stratified random sub-sampling is used. The split ratio is 80:20, with 800 and 200 images randomly divided into training and testing sets, respectively. In order to use the best n-gram features in MSR, five n-gram range sets are tested, from 1-gram up to 5-gram. Two types of weighting schemes are used to test the classification performance, namely TF, and TF-IDF. Weighting schemes with better classification performance will be selected to identify relevant features using Information Gain(IG) feature selection. Four threshold values of 20%, 40%, 60%, and 80% are used to select a subset of features that have a significant impact on the model's performance. This threshold value represents a cutoff point for feature importance or relevance. Features with values above this threshold are selected, while features with values below the threshold are discarded. The SVM classifier is used to classify the images.

4.3 Results on Dredze Dataset

According to preliminary training set experiments, the TF weighting performs best at the 3-gram setting compared to the TF-IDF weighting, and these accuracy results are consistent with the accuracy of the test set, as shown in Table 1. The best accuracy performance was attained at the 3-gram with TF weighting, which recorded 92.70%. Hence, the 3-gram features with TF weighting were chosen to be used in the MSR method. Experiment on 4-gram and 5-gram cannot be carried out because these two gram sizes have a very large size of feature descriptors and require a large memory to be implemented.

Table 1: The Average Classification Accuracy (Mean and Standard Deviation) on Dredze Dataset for 1-gram, 2-gram and 3-gram

	TF	TF-IDF
1-gram	83.38 ± 2.05	82.19 ± 2.11
2-gram	91.10 ± 2.00	90.83 ± 2.12
3-gram	**92.70 ± 1.82**	92.55 ± 2.00

The feature size for 3-grams is 262,157 features. This large number of feature descriptors can increase the risk of overfitting and reduce the generalization ability of the model. To address this issue, a smaller number of feature descriptors is used in the experiment, which is 1,000 feature descriptors. Table 2 shows the average classification accuracy in percentage (%) and standard deviation of the Dredze dataset using the MSR method.

Table 2: The Average Classification Accuracy (Mean and Standard Deviation) on Dredze Dataset using MSR

	Accuracy
Level 0	92.65 ± 1.86
Level 1	93.58 ± 1.64
Level 2	93.85 ± 1.78
Level 3	94.04 ± 1.73

Due to the accuracy of TF being higher than TF-IDF for level 0, the feature vector with TF weighting scheme is used in the IG feature selection. Table 2 shows that the best average classification accuracy for each level increases according to the size of the level. At level 0, the threshold value of 80% results in the training set's highest classification performance, and the test set's evaluation of the model with this threshold value yields the results shown in Table 2, which is 92.65%. The result of the training set with a threshold value of 80% outperformed the results of the training set with all features taken into account. At level 1, the threshold value of 80% still gives the best classification performance on the training set, resulting in 93.58% on the test set. Next, at level 2, the threshold value of 80% continues to deliver the training set's highest classification performance, and the model's evaluation of the test set using this threshold value results in a classification accuracy of 93.85%. Finally, level 3 recorded 94.04% accuracy with a threshold value of 60%.

Table 3: The Average Classification Accuracy (Mean and Standard Deviation) on Dredze Dataset using Ensemble Method

Mean	Product	Weighted Mean
94.06 ± 1.77	94.04 ± 1.77	**94.40 ± 1.67**

Table 3 shows the average classification accuracy of the ensemble method. The model for each level will be selected as input to the ensemble of MSR. Class probabilities for the model at level 0, level 1, level 2, and level 3 are used as input to the ensemble of MSR. The three combination methods used to measure classification accuracy are mean, product, and weighted mean. Classification accuracy results of the ensemble method using weighted mean outperform the best classification accuracy result from the MSR method (94.04% at level 3). The weight values for the weighted mean are obtained through PSO. The classification accuracy for the ensemble of MSR using a weighted mean is 94.40%. The weights for level 0, level 1, level 2, and level 3 are 0.00%, 60.00%, 72.98%, and 72.98%, respectively. However, average and product combination methods give results that are almost the same as the best results for MSR, with 94.06% and 94.04%, respectively. Figure 4 shows the graph results on Dredze dataset. From the graph, it is found that the classification accuracy for each level increases according to the size of the level. Meanwhile, the ensemble of MSR using weighted mean improved the accuracy of MSR and outperformed all four MSR models for Dredze dataset.

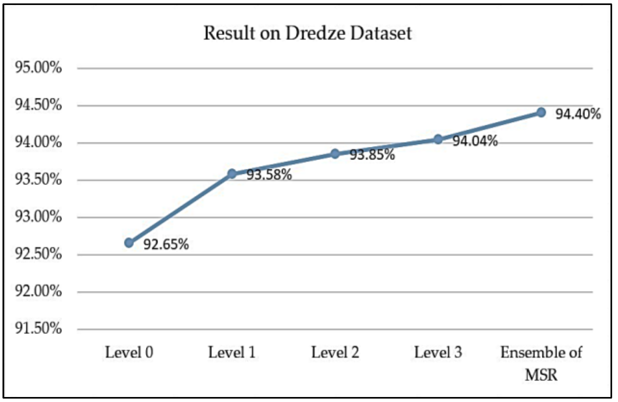

Figure 4: Classification accuracy result on Dredze dataset.

4.4 Results on SpamArchive Dataset

Based on initial experiments, similar to the Dredze dataset, the TF weighting performs best at the 3-gram setting compared to the TF-IDF weighting for SpamArchive. Table 4 shows the accuracy of the test set where the best accuracy performance was attained at the 3-gram with TF weighting, which recorded 94.35%.

Table 4: The Average Classification Accuracy (Mean and Standard Deviation) on SpamArchive Dataset for 1-gram, 2-gram and 3-gram

	TF	TF-IDF
1-gram	83.19 ± 2.13	83.06 ± 2.32
2-gram	90.23 ± 2.03	89.14 ± 2.00
3-gram	**94.35 ± 1.53**	91.28 ± 2.09

The 3-gram features and 1,000 feature descriptors were used for SpamArchive. Table 5 shows the average classification accuracy(%) and standard deviation of the SpamArchive dataset using the MSR method. Similar to the classification results on Dredze, the average classification accuracy increases with level size. The classification performance increased by 0.42% from level 0 to level 1, from 94.35% to 94.77%, and an increase of 0.16% from level 1 to level 2, from 94.77% to 94.93%. However, the classification performance decreased by 0.29%, from 94.93% to 94.64%. The performance of level 3 is slightly worse than the performance at level 1 and level 2 because it is possible that the feature descriptor size is highly dimensional, and local features at level 3 cannot provide good classification performance. Level 1, level 2, and level 3 use a feature selection threshold value of 80%, where the training set provides the best performance. For level 0, the best training set performance is without feature selection.

Table 5: The Average Classification Accuracy (Mean and Standard Deviation) on SpamArchive Dataset using MSR

	Accuracy
Level 0	94.35 ± 1.53
Level 1	94.77 ± 1.52
Level 2	94.93 ± 1.57
Level 3	94.64 ± 1.64

Table 6 shows the average classification accuracy of the ensemble method for SpamArchive. Class probabilities for the model at level 0, level 1, level 2, and level 3 are used as input to the ensemble of MSR. All classification accuracy results of the ensemble method outperform the best classification accuracy result from the MSR method (94.93% at level 2). The accuracy results for the weighted mean, mean, and product produced nearly identical results, with an accuracy of 95.15%, 95.13%, and 95.11%, respectively. The weights generated by PSO for the weighted mean for level 0, level 1, level 2, and level 3 are 41.00%, 40.82%, 44.02%, and 95.07%. Figure 5 shows the graph results on SpamArchive dataset. From the graph, it is found that the classification accuracy for each level increases according to the size of the level until level 2, and started decreased at level 3. Meanwhile, the ensemble of MSR using weighted mean improved the accuracy of MSR and outperformed all four MSR models for SpamArchive dataset.

Table 6: The Average Classification Accuracy (Mean and Standard Deviation) on SpamArchive Dataset using Ensemble Method

Mean	Product	Weighted Mean
95.13 ± 1.48	95.11 ± 1.48	**95.15 ± 1.49**

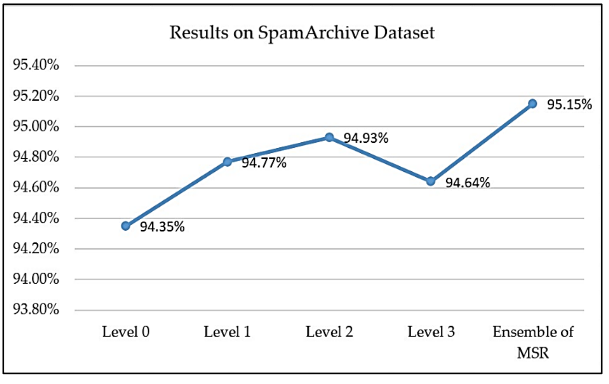

Figure 5: Classification accuracy result on SpamArchive dataset.

5.0 CONCLUSION

Two methods were proposed, MSR and the ensemble of MSR. MSR divides Base64 codes into smaller partitions to incorporate spatial information between features. The feature vector is produced by computing histograms of n-gram Base64 codes within each of these partitions. Results from the experiment show that MSR can be applied to the text where the accuracy of each level increases as the level progresses from L0 to L3. The accuracy results of MSR can be further improved where the class probabilities from all four models are used as input to the ensemble of MSR. This study utilized three combination methods, and the weighted mean gives the highest classification accuracy for both datasets.

REFERENCES

[1] A. Annadatha and M. Stamp, "Image spam analysis and detection," *Journal Computer Virology. Hacking Techniques*, vol. 14, no. 1, pp. 39–52, Feb. 2018. doi: 10.1007/s11416-016-0287-x.

[2] A. Barushka and P. Hajek, "Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks," *Neural Computing Applications*, vol. 32, no. 9, pp. 4239–4257, May 2020, doi: 10.1007/s00521-019-04331-5.

[3] G. Fumera, I. Pillai, and F. Roli, "Spam Filtering Based On The Analysis Of Text Information Embedded Into Images," *Journal Machine Learning Research*, vol. 7, pp. 2699–2720, 2006.

[4] X. Bin, L. Ruiguang, L. Yashu, Y. Hanbing, L. Siyuan, and Z. Honggang, "Filtering Chinese image spam using Pseudo-OCR," *Chinese Journal Electron*, vol. 24, no. 1, pp. 134–139, 2015, doi: 10.1049/cje.2015.01.022.

[5] Y. K. Zamil, S. A. Ali, and M. A. Naser, "Spam image email filtering using K-NN and SVM" *International Journal Electrical and Computer Engineering*, vol. 9, no. 1, p. 245, 2019, doi: 10.11591/ijece.v9i1.pp245-254.

[6] N. N. Abuzaid and H. Z. Abuhammad, "Image SPAM Detection Using ML and DL Techniques," *International Journal Advance Soft Computing Appication*, vol. 14, no. 1, pp. 226–243, 2022, doi: 10.15849/IJASCA.220328.15.

[7] H. Ghizlane, R. Jamal, M. A. Mahraz, Y. Ali, and T. Hamid, "Spam image detection based on convolutional block attention module," 2022 Int. Conf. Intell. Syst. Comput. Vision, ISCV *2022*, pp. 0–3, 2022, doi: 10.1109/ISCV54655.2022.9806065.

[8] M. Dredze, R. Gevaryahu, and A. Elias-Bachrach, "Learning Fast Classifiers for Image Spam," in Proceedings of the Fourth Conference on Email and Anti-Spam (CEAS' 07), 2007, pp. 487–493.

[9] C. Xu, Y. Chen, and K. Chiew, "An Approach to Image Spam Filtering Based on Base64 Encoding and N-Gram Feature Extraction," 2010 22nd IEEE Int. Conf. Tools with Artif. Intell., pp. 171–177, Oct. 2010, doi: 10.1109/ICTAI.2010.31.

[10] C. Leng, H. Zhang, B. Li, G. Cai, Z. Pei, and L. He, "Local Feature Descriptor for Image Matching: A Survey," *IEEE Access*, vol. 7, pp. 6424–6434, 2019, doi: 10.1109/ACCESS.2018.2888856.

[11] J. Revaud, J. Almazan, R. Rezende, and C. De Souza, "Learning with average precision: Training image retrieval with a listwise loss," Proc. IEEE Int. Conf. Comput. Vis., vol. 2019-October, pp. 5106–5115, 2019, doi: 10.1109/ICCV.2019.00521.

[12] F. Radenovic, G. Tolias, and O. Chum, "Fine-Tuning CNN Image Retrieval with No Human Annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, 2019, doi: 10.1109/TPAMI.2018.2846566.

[13] Z. Hu and A. G. Bors, "Expressive Local Feature Match for Image Search," Proc. - Int. Conf. Pattern Recognit., vol. 2022-August, pp. 1386–1392, 2022, doi: 10.1109/ICPR56361.2022.9956438.

[14] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, vol. 2, pp. 2169–2178, doi: 10.1109/CVPR.2006.68.

[15] P. Karmakar, S. W. E. I. Teng, G. Lu, and D. Zhang, "An Enhancement to the Spatial Pyramid Matching for Image Classification and Retrieval," *IEEE Access*, vol. 8, pp. 22463–22472, 2020, doi: 10.1109/ACCESS.2020.2969783.

[16] B.-D. Liu, J. Meng, W.-Y. Xie, S. Shao, Y. Li, and Y. Wang, "Weighted Spatial Pyramid Matching Collaborative Representation for Remote-Sensing-Image Scene Classification," *Remote Sensors*, vol. 11, p. 518, 2019, doi: 10.3390/rs11050518.

[17] A. Ashiquzzaman, H. Lee, K. Kim, H.-Y. Kim, J. Park, and J. Kim, "Compact Spatial Pyramid Pooling Deep Convolutional Neural Network Based Hand Gestures Decoder", *Applied Sciences*, vol. 10, no. 21, p. 7898, Nov. 2020, doi: 10.3390/app10217898.

[18] A. Abdullah, R. C. Veltkamp, and M. A. Wiering, "Spatial Pyramids and Two-layer Stacking SVM Classifiers for Image Categorization: A Comparative Study," International Joint Conference on Neural Networks, pp. 5–12, 2009.

[19] M.-L. Zhang and Z.-H. Zhou, "Exploiting Unlabeled Data to Enhance Ensemble Diversity," *Data Mining and Knowledge Discovery*, vol. 26, no. 1, pp. 98–129, Sep. 2009, doi: 10.1007/s10618-011-0243-9.

[20] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, no. August, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[21] L. Rokach, "Chapter 45 Ensemble Methods For Classifiers," 1999.

[22] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 421, pp. 123–140, 1996, doi: 10.1007/BF00058655.