Features Assimilation via Three-Stream Deep Networks for Spam Attack Detection from Images

Ashish Yadav, Shubham Kumar, Anirudh Chaudhary, Devansh Goyal,
Amanjit Singh, Samarth Roday, and Tushar Sandhan

Indian Institute of Technology, Kanpur, India
{ashisy20, kshubham20, anirudhc20, devanshg20, samanjit20, samarthr20, sandhan}@iitk.ac.in

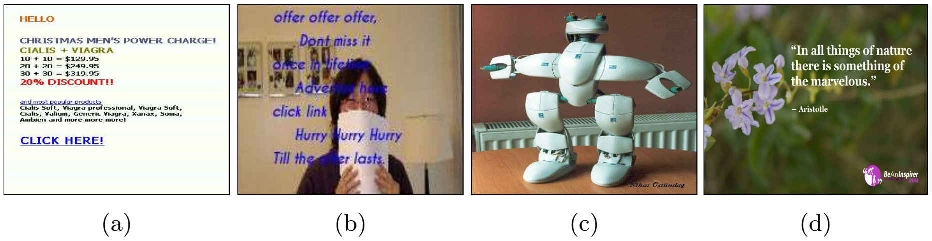
Abstract. Spam filters typically use optical character recognition (OCR) for extracting the text from images. These days spammers have circumvented optical scanning by fracturing the text within the images thereby improving their attacks and finally reaching to the users. This paper proposes a three-stream deep learning-based model which uses Convolutional Neural Networks (CNN), Transfer Learning, SIFT and HOG features via hybrid fusion framework. Transfer learning alone can only achieve an accuracy of 95% but our hybrid model shows improved performance and obtains an accuracy of 96%, eclipsing the existing techniques. We have created our dataset of challenging HAM images which will be publicly available. On our challenging dataset as well, the proposed method outperforms other existing methods for effectively detecting the spam attacks targeted via images.

Keywords: · Convolutional neural networks · Transfer learning· Spam image · Hybrid 3-stream model· Bag of words model.

1 Introduction

E-Mail has become a ubiquitous communication medium and widely popular nowadays. According to the report released by the Radicati group [1], as of April 2022, there are 4.26 billion email users worldwide, which is approximately fifty percent of the world population. However, the effectiveness of email has often been reduced due to compromised security by spam attacks via emails. Spam emails, also known as junk emails, are uninvited email messages that are typically delivered to a large number of recipients. Every day, hackers, invaders, and attackers seek to exploit consumers by sending several unsolicited emails containing unwanted information. To combat this problem, a number of Machine Learning (ML)-based spam detectors were created. Initially, spam from e-mail was in the form of text. ML models such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM) and Naive Bayes (NB), among others, are used to filter email spam based on textual content and have achieved up to 95 percent accuracy [2].

With the technology enhancement of spam detection, the attackers are always finding a new way to spam the users even with the multiple layers of security


HELLO
CHRISTMAS MEN'S POWER CHARGE!
CIALIS + VIAGRA
10 + 10 = $129.95
20 + 20 = $249.95
30 + 30 = $319.95
20% DISCOUNT!!
and most popular products
Cialis Soft, Viagra professional, Viagra Soft, Cialis, Valium, Generic Viagra, Xanax, Soma, Ambien and more more more!
CLICK HERE!
offer offer offer,
Dont miss it
click link
Hurry Hurry
Till the
r lasts.
"In all things of nature there is something of the marvelous."
– Aristotle
BeAnInspirer

(a) (b) (c) (d)

Fig. 1: Spam text in form of image is shown in (a), while in (b) spam text is embedded over natural image to make it more challenging. Ham image is shown in (c), while it is more challenging to recognize (d) as a ham image because useful texts are embedded over the natural image.

mechanisms. They find the bugs, vulnerabilities and exploit them by actively improving their spam attacks on a day-to-day basis. In recent times, they have found a new way of sending spam emails through embedding all the information inside a single image. Whereas the real word normal image containing genuine information is called as Image Ham.

Email spam in which the spam text is embedded in an image is known as 'Image spam', as shown in Fig. 1a. To improve the spam attack further, spammers have made image spam more challenging by embedding spam text over natural images like Fig. 1b. Now the problem is that, the distinction between spam and ham images became so narrow that even human can not separate them easily. In challenging ham images, useful texts are embedded over the natural images Fig. 1d. With dire need of strong Cybersecurity in this digital age, the image spam detection is very crucial problem to be solved.

In the initial stages, the text is embedded within an email in the form of HTML, and to counter that researchers have started using optical character recognition(OCR) techniques [3]. OCR is a technology that recognizes text in digital images. But later spammers started using captcha-based techniques to obfuscate the text in the images which was difficult to read by OCR algorithms. This problem motivated the researchers to propose several machine learning and deep learning based algorithms for effective classification of image spams [4]. Numerous researchers have developed approaches based on deep learning for diverse cyber security applications [5], including malicious domain discovery [8], malware detection [6], [7], intrusion detection [9] etc. Transfer learning and many pre-trained CNN models, such as Xception, VGG19 etc., have been utilised in the work presented in [10]. Transfer learning alone is not sufficient to protect against sophisticated spam image attacks. In addition to the framework for transfer learning, we suggest fusing a multi-feature hybrid three-stream deep convolutional neural network for efficient spam detection.

The main contribution of this work is the following: Firstly, four models SIFT feature CNN, SIFT image CNN, HOG feature CNN, and Hog image CNN models are proposed and their effectiveness is analyzed for image spam detection.

Secondly, the ability of the transfer learning techniques are studied by utilizing pre-trained CNN model VGG19. Lastly, the features extracted by HOG [11], SIFT [12], and VGG19 [13] are combined using a 3-stream hybrid CNN model and used for the detection of spam images. The remaining sections of this work are organized as follows; Section 2 presents the literature review, Section 3 contains the proposed model architecture and Section 4 presents the dataset, experiments, and results. Finally, the Section 5 concludes our work.

2 Literature Review

The recent state-of-the-art methods employ deep neural networks for spam image classification. Image spam classification based on CNN [14] method proposes a novel deep architecture, where a linear support vector machine (LSVM) is used in the output layer, and parameters are tuned by minimizing a loss which is marginal in nature correspondingly. They trained the network with hinge-loss instead of softmax-loss i.e. they placed the classic softmax layer with L2-SVM while optimizing parameters via backpropagation from the last SVM layer focusing on assessing feature representation performance on the spam classification task. The network compromises of five convolutional layers and three fully-connected layers. Images are rescaled to 256×256 and a crop of 227×227 is passed to the network. Finally, the output obtained from the last layer which was fully connected is passed to a SVM layer to train the network for classification. Backpropagation of gradients from the top linear SVM layer is used to train the lower layer weights. Unlike the previous methods the method in [15] uses the multi-estimating points to enhance the standard moment estimation and ADAM [16] a stochastic gradient optimization algorithm.It also proposes WDSP-net combined with SPR and ADAM algorithms. The WDSP-net achieves very high accuracy in image spam recognition.

In [10], two DCNN and hybrid models are studied on 3 different datasets for image spam classification. Balanced class weights is used to study the effects of cost-sensitive learning on model accuracy and pre-trained CNN architectures like Xception [18], VGG19 [17], etc. are used to study effects of transfer learning. Some parts of CNN models are trained on several combinations of Dredze Image-Spam dataset for 100 epochs in various input sizes. Then these models are trained and tested on the ISH dataset for improving the performance. Hybrid models are also employed which extracted the features from the last hidden dense layer of the base model to enhance the performance.

Color Model Based CNN is proposed in [19] for image spam classification. It consists of two steps- image preprocessing and CNN. Their CNN model for image spam detection is tiny for speed improvement and it has an input layer, three convolutional layers, three max-pooling layers, a flatten layer, a drop-out unit, and two dense fully connected layers. Their color space analysis, RGB experiments did not obtain as good accuracy as YCbCr, YUV, and XYZ. According to [19] the XYZ color model achieved the highest accuracy among all color models. In comparison to other related works, their XYZ model-based CNN was able to increase the accuracy up to 98.4% on ISH dataset where previous best performance reported was 98%. From the results, it can be said that different color models obtained different accuracy.

Keeping this in mind, we have performed extensive experiments on different color models. Three machine learning techniques were tested in [20] SVM and two techniques based on neural-net, CNN and multilayer perceptrons (MLP). [20] also studied the features based on Canny images, row images [21], and their combination. Their detailed experiments were based on three datasets established the effectiveness of the proposed approaches, and concluded that a SVM model achieved high accuracy on a public image spam dataset, a CNN technique performed better on challenging image spam dataset. On the ISH dataset all three techniques described in this paper performed well, with SVM achieving an accuracy of 98.72 while a CNN achieving even more at 99.02, while an MLP showed an accuracy of 95.57. For challenging dataset 1, CNN surpassed other methods with an accuracy of 83.13, while none of the techniques can achieve an accuracy greater than 71.83 on the more challenging challenge dataset 2. Moreover these challenging datasets are not publicly available. Nevertheless it shows how dramatically the best performing model on easier spam dataset can show reduced performance on just slightly challenging spam image dataset.

The above discussion has shown some important feature extraction techniques that have been used to extract important features from an image. Some methods are manual i.e employ image processing techniques like HOG and SIFT while some methods are based on learning feature extraction from training like CNN. These techniques have been used in many previous works. Now let's analyze some of these techniques as they form a building block for our three-stream network framework:

Scale Invariant Feature Transform (SIFT): SIFT [12] is a image processing algorithm which describes and detects local features in images. For each image, SIFT is applied to extract the key-points from the image. After locating the key-points, the magnitude and direction of the gradient are calculated using neighboring pixels of the key point. To identify the dominant directions, the gradient histogram is formed. The number of key-points obtained is different for different images. So, it cannot be used directly to give input to a CNN model. So for this we perform, and then for clustering of key-points we employed the Bag of Words model [22].

Bag of Words (BOW): The BOW model [22] is used for vectorization of text, i.e it turns random text into fixed-length vectors by counting the number of times each word appears. This is done because machine learning algorithms is incompatible to work with the raw text directly, the text needed to be converted into numbers. Similar concept is used for obtaining visual words inside an image and a dictionary is declared to hold the bag of words and tokenize each feature into words. Now for each word in an image, it is checked that if the word exists in the dictionary. If it does, then count is incremented by 1. If it doesn't, it is added to the dictionary and count is set as 1.

Histogram-Oriented Gradients (HOG): The HOG [11] is a powerful element descriptor utilized broadly in object identification and recognition. While the HOG features are essentially utilized for object identification, object recognition can also be performed using the extracted features. The fundamental difference among HOG and other feature descriptors, for example, shape contexts, SIFT, and histograms which are edge oriented, is that dense uniformly dispersed grids are utilized for the calculation of HOG and nearby contrast normalization is utilized for improvement of accuracy. The HOG descriptor is photometric and geometric transformation invariant. These characteristics enabled us to investigate the assimilation of two different features viz. SIFT and HOG for detecting spam pictures.

Support Vector Machines: The following details will explain the key concepts of SVM as mentioned in [23]:

Separating Hyperplane - During training, SVM tries to find a hyperplane that separate different classes by acting as a decision boundary. Obviously, such a hyperplane need not exist, which leads us to think about spaces which are higher dimensional.

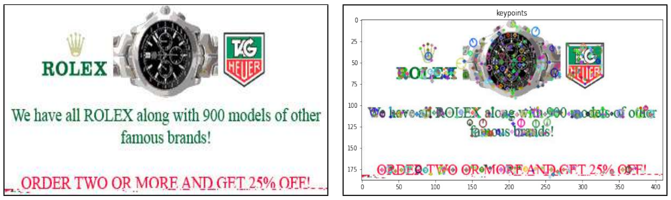
Maximize the margin - Classes are separated using a hyperplane,in that case there will be endless numbers of such hyperplanes. In SVM, a hyperplane is picked that enable it to maximizes the margin. Here the least distance between the class of data and hyperplanes are referred as margin. We have employed SVM at the final stage for the classification.

3 Proposed Method

3.1 Pre-processing

Dataset is randomly divided into train and test dataset for training and testing purpose. The training dataset contains 70 percent images of the whole dataset i.e 2690 images while the test dataset contains 30 percent of the total images i.e 1154. Each image from the train and test dataset is converted from RGB to grayscale. The train and test grayscale images are used to extract SIFT and HOG features.

SIFT Feature: The SIFT descriptor [12] is obtained by dividing an image into 4x4 squares. For each of these sixteen squares, a vector of length eight is obtained. By merging all the vectors, a vector of size 128 is formed for each key point. To utilize the generated key-point descriptors in classification, a fixed size vector is required. For this purpose, bag of words model is employed which uses K-means


ROLEX
TAG HEUER
We have all ROLEX along with 900 models of other famous brands!
ORDER TWO OR MORE AND GET 25% OFF!
keypoints
0
25
50
75
100
125
150
175
0
50
100
150
200
250
300
350
400

(a) (b)

Fig. 2: Embedded key-points are shown in (b) of spam Image shown in (a) using SIFT features.

to cluster the descriptors into a group. Then, a bag of key-points is created by calculating descriptors number that are enclosed in every cluster. The resulting feature vector has a definite size.

HOG feature: In HOG [11], the histogram of oriented gradients is used as feature. Gradients which are calculated in x and y directions are features that represent complex shapes like edges and corners. The gradient's direction represents the directional change in intensity of pixels in an image, whereas the amount of the change is represented by pixels.

$$\nabla f = \begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} \frac{\delta f}{\delta x} \\ \frac{\delta f}{\delta y} \end{bmatrix} \tag{1}$$

Here, the derivative of the image with respect to x and y is given by $\frac{\delta f}{\delta x}$ and $\frac{\delta f}{\delta y}$ respectively. The derivative can further be calculated as shown in 2 and 3:

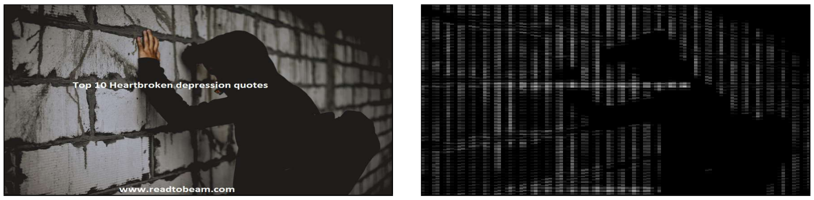
$$f_x(x) = \frac{\delta f}{\delta x} = f(x+1) - f(x-1) \tag{2}$$

$$f_y(y) = \frac{\delta f}{\delta y} = f(y+1) - f(y-1) \tag{3}$$

After calculation of gradients, the magnitude and direction of the gradients can be obtained using the equation 4:

$$f = \sqrt{f_x(x)^2 + f_y(y)^2}, \quad \theta = \arctan \frac{f_x(x)}{f_y(y)} \tag{4}$$

In cases of corners and edges there are sudden large changes in intensity making the magnitude of gradient to be large. In smooth regions there are no sudden changes in intensity as a result the gradient magnitude is zero. Thus, while calculating the gradients various redundant information in the in image background is eliminated [11]. The Normal ham Image is shown in Fig. 3a while Fig. 3b shows the HOG image extracted from the given ham image.



(a) (b)

Fig. 3: A ham image is shown in (a) and its corresponding transformed HOG feature image is shown in (b), which becomes an input for HOG-CNN.

```
Algorithm 1: Image Spam Classification
input        : A set of images obtained from emails
output       : Labels $y_1 \dots y_N$
               (0 for Ham and 1 for Spam)
preprocessing: Duplicates are removed and images are resized to desired
               sizes as mentioned in Table1
for Every image do
    Convert from RGB to GRAY
    for GRAY image do
        Extract SIFT descriptors
        Apply Bag of Words model on SIFT descriptor to capture all
        key-points
        Extract HOG image
    Pass RGB image to VGG19 feature extractor containing VGG19 to
      extract optimal feature vector $v_i^1$
    Compute $d_i^1 = Denselayer(v_i^1)$
    Pass SIFT descriptors to SIFT CNN to extract optimal feature vector $v_i^2$
    Compute $d_i^2 = Denselayer(v_i^2)$
    Pass HOG images to HOG CNN to extract optimal feature vector $v_i^3$
    Compute $d_i^3 = Denselayer(v_i^3)$
    Concatenate $D_i = [d_i^1; d_i^2; d_i^3]$
    Compute $D_i^1 = Denselayer(D_i)$
    Calculate $y_i = Sigmoid(D_i^1)$
```

Network Architecture: We propose a three-stream network as shown in Fig. 4 using pre-trained VGG-19 and custom CNN layers with modifications and finally classification is done using SVM as a classifier. Our network consists of three CNN models applied to the transformed data from the SIFT, HOG, and normal image respectively and the features from three streams are concatenated to form the final model. The description of each individual model is given below as well as summarized in Algorithm 1.

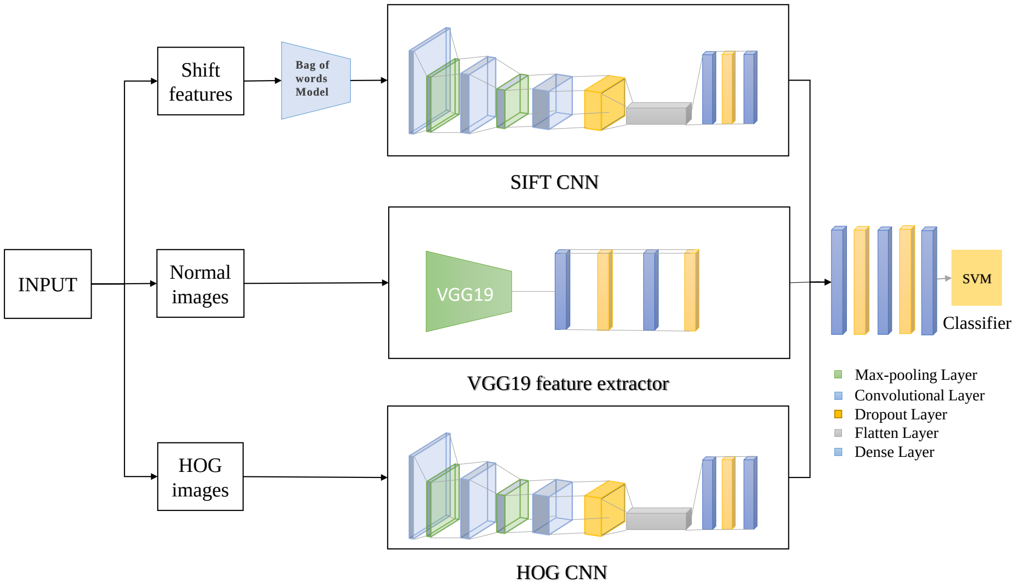


Fig. 4: Proposed Pipeline of Three-Stream Hybrid Model.

Top-stream (SIFT CNN: Input Size $50\times128\times1$*)* : It takes resized SIFT descriptors as input and applies convolutional layers for feature extraction. It also has 3 Convolutional layers and 2 Max-pooling layers for feature extraction and downsampling of the feature space respectively. Finally, the features are flattened and passed through a dense layer with appropriate dropouts. Top and bottom stream's CNN architecture is kept similar for symmetry and throughput improvement but models are having different filter weights. The features obtained from all these models are concatenated and passed through a dense layer with an intermediate dropout layer to avoid overfitting the model. The output from the dense layer is passed to SVM for the final classification.

Mid-stream (VGG19 transfer learning features :Input size $156\times156\times3$*)* : This model is based on the concept of transfer learning. It takes the image as input and passes through the pre-trained VGG-19 model trained on the 'imagenet' dataset. The features extracted from VGG-19 are flattened and passed through dense layers. Appropriate dropouts have been applied between dense layers.

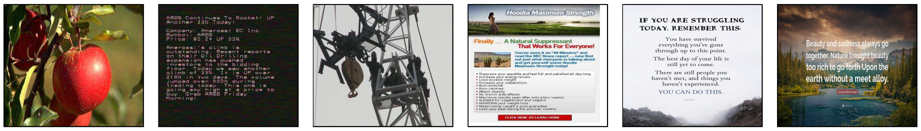
Bottom-stream (HOG CNN :Input size $156\times156\times1$*)* : It takes grayscale HOG images as input and applies convolutional layers to extract features from the input. It has 3 Convolutional layers and 2 Max-pooling layers for feature extraction and downsampling of the feature space respectively. Finally, the features are flattened and passed through a dense layer with appropriate dropouts. The model summary of the entire unified framework is discussed in Table 1.

Table 1: Model summary describing each layer with output shape and number of parameters.

Layer Type	Outputs	Parameters	Layer Type	Outputs	Parameters
Input Layer	(-,156,156,1)	0	Flatten	(-,36992)	0
Input Layer	(-,50,128,31)	0	Flatten	(-,49152)	0
Conv2D	(-,156,156,32)	320	Dense	(-,512)	4194816
Conv2D	(-,50,128,32)	320	Dense	(-,512)	18940416
Maxpooling2D	(-,52,52,32)	0	Dense	(-,512)	25166336
Maxpooling2D	(-,25,64,32)	0	Droupout	(-,512)	0
Conv2D	(-,52,52,64)	18496	Droupout	(-,512)	0
Conv2D	(-,25,64,64)	18496	Droupout	(-,512)	0
Maxpooling2D	(-,17,17,64)	0	Dense	(-,128)	65664
Maxpooling2D	(-,12,32,64)	0	Dense	(-,128)	65664
Input Layer	(-,156,156,3)	0	Dense	(-,128)	65664
Conv2D	(-,17,17,128)	73856	Concatenate	(-,384)	0
Conv2D	(-,12,32,128)	11	Dense	(-,384)	147840
VGG19(Funtional)	(-,4,4,512)	20024384	Droupout	(-,384)	0
Droupout	(-,17,17,128)	0	Dense	(-,128)	48280
Droupout	(-,12,32,128)	0	Droupout	(-,128)	0
Flatten	(-,8192)	0	Dense	(-,1)	129
Total Parameter count = 68,905,537					

4 Experiments

All of our experiments have been performed using NVDIA RTX 3660 GPU with 12GB of RAM. We have used Python, OpenCV, and the Tenserflow library. We have experimented with different image color spaces like BGR, YCbCr, LAB, etc. We tested various techniques, other features amd classifiers like pure HOG features, K-means clustering and SVM with our dataset. The following is the description of the datasets used in our research and Fig. 5 shows some of the samples images.


ARSS Continues To Rocket! UP Another 33% Today!
Company: Amerossi EC Inc.
Symbol: ARSS
Price: $0.24 UP 33%
Hoodia Maximum Strength
Finally ... A Natural Suppressant That Works For Everyone!
CLICK HERE TO LEARN MORE
IF YOU ARE STRUGGLING TODAY, REMEMBER THIS:
You have survived everything you've gone through up to this point.
The best day of your life is still yet to come.
There are still people you haven't met, and things you haven't experienced.
YOU CAN DO THIS.
Beauty and sadness always go together. Nature thought beauty too rich to go forth Upon the earth without a meet alloy.

(a) (b) (c) (d) (e) (f)

Fig. 5: Ham image in (a) and spam text embedded image in (b) are from Dredze Image Spam dataset. Ham image in (c) and spam text embedded image in (d) are from ISH dataset. Ham image (e) and (f) with embeddded useful text to make it more challenging to be recognized as Ham image are from our dataset.

4.1 Datasets

Dredze Image Spam Dataset: The dataset [24] contains images in 3 sets. Personal Spam has 3,298 images in total out of which the number of unique images are 1,274 . Personal Ham has 2,021 images out of which the number of unique images are 1,517. And, the Spam Archive has files of various formats like PNG, JPEG, GIF etc, in total it has 16,028 files in which the number of unique images are 3,039.

Image Spam Hunter Dataset (ISH): The ISH dataset [25] contains both ham and spam images collected from original emails and images are in in JPEG format. There are 929 spam and 810 ham images in total. The number of unique ham and spam images found in the dataset after processing is 810 and 879 respectively.

Our HAM Dataset: This dataset is created by web crawling to get more challenging HAM images with embedded text, challenging images will allow us to train our model more efficiently which will increase the efficiency of the model in differentiating between challenging HAM and spam images with embedded text. There are 2533 ham images in this dataset. Our dataset will be publicly available.

4.2 Multidataset Unification

The datasets that are used in this work contain several duplicate images and corrupted files. At first, the corrupted files are deleted and then to avoid duplicated files, hashing is done to convert each image into a hash and stored in hashlist. In this way, whenever a duplicate image is encountered, its hash will be matched with the images present in the hashlist. If the match is confirmed, then the image will be skipped. Lastly , all unique images are resized into (156,156,3) shapes.

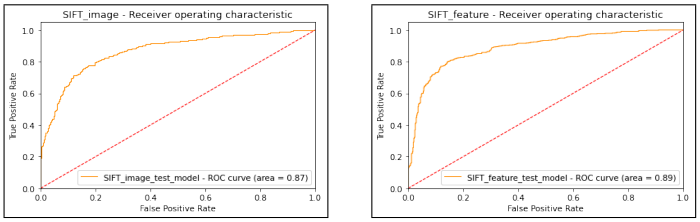
4.3 Ablation Study

Effectivness of the Top Stream: (Only top-stream from Fig. 4 is used for this study) Grayscale dataset was obtained from original dataset. Then SIFT feature vector was constructed after resizing extracted SIFT descriptor to (50,128,1). Resizing is required as we obtain a different number of key-points for a different image.

SIFT key-points were plotted to the corresponding image and stored to check the accuracy of the model on both the SIFT image and feature. SIFT feature and SIFT image both dataset was divided into 30 percent test and 70 percent train. Test model was trained on both SIFT feature and SIFT image and tested its accuracy respectively. The test model shows higher accuracy 82 percent accuracy on SIFT feature for predicting spam and ham image than on SIFT image which has an accuracy of 80 percent as given in Table 2 and Fig. 6.

Table 2: Percentage accuracy of test CNN on SIFT image vs SIFT feature.

Model	Accuracy
SIFT Image + CNN	80
SIFT Descriptor +CNN	82



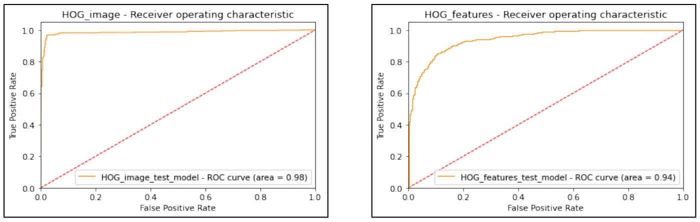
(a) (b)

Fig. 6: ROC of SIFT-CNN with (a) SIFT image and (b) SIFT feature as input.

Effectivness of the Bottom Stream: (Only bottom-stream from Fig. 4 is used here) Grayscale was obtained from original dataset. Both the HOG feature vector and HOG image was obtained. Obtained size of the HOG feature array was resized as the required input size. Both the HOG feature array and HOG image array are divided into 30 percent test and 70 percent train. CNN model is trained and tested on both datasets separately. Obtained accuracy was 80 percent for the HOG features and 90 percent for HOG images. So clearly, the HOG image performed better than the HOG feature for spam image classification. This accuracy is given in Table 3 and Fig. 7.

Table 3: Percentage accuracy of test CNN with HOG image vs HOG feature.

Model	Accuracy (%)
HOG Image + CNN	90
HOG Feature + CNN	80



(a) (b)

Fig. 7: ROC of HOG-CNN with (a) HOG image and (b) HOG feature as input.

SVM with Linear and RBF kernel: Our test model is tested on SVM as a classifier. Linear and RBF kernels of SVM are used to check the model accuracy one by one. In Table 4 LSVM refers to SVM with linear kernel while RSVM refers to SVM with RBF as kernel. SVM with RBF as kernel gave higher accuracy as compared to linear kernel as shown in Table 4.

Table 4: Percentage accuracy of SVM using linear and RBF kernels.

Classifier	Accuracy (%)
LSVM (Linear kernel)	95
RSVM (RBF kernel)	96

Color Space Analysis: Pre-trained test model was tested after transforming our image dataset from RGB to other image types like LAB, HSV, YCbCr, etc. Pixel value of GRAY images captures the intensity of light according to a particular weighted combination of frequencies. The HLS color space module converts the image into a hue, saturation, and lightness components. The Hue is the color of the image, the Saturation is the pureness of the hue, and the lightness is the strength of the hue. LAB contains a mix of one channel with no color (L), plus two channels that have no contrast but with a dual color combination (A+B). In YCbCr, Y is the luma, the brightest component of the color.It represents the brightness of the color. Cr and Cb are the red and blue component relative to the green component respectively. Accuracy of different color models are given in Table 5.

Table 5: Performance of test CNN on different Color spaces.

Model	**Precision**	**Recall**	**F1-Score**	**Support**
BGR2GRAY	0.86	0.82	0.82	1147
BRG2HLS	0.69	0.63	0.59	1147
BGR2LAB	0.73	0.58	0.49	1147
BGR2YCbCr	0.23	0.38	0.28	1142

4.4 Quantitative Analysis Comparison with Other Methods

After pre-processing, the whole dataset containing 1937 spam images and 1907 ham images is divided into 30 percent test and 70 percent train. SIFT and HOG features are extracted from the train and test dataset. Bag-of-words model using K-means clustering is applied on SIFT descriptors. Finally, all three train datasets are used to train our model upto 100 epochs with batch size of 32. A learning rate of 0.001 is applied with Adam as optimizer and cross-entropy as loss function. Testing of our proposed method gave an accuracy of 96 percent as shown in Table 6. Performance of various transfer learning methods like VGG19, Xception, ResNet251 and DenseNet201 are studied on our dataset. ResNet251 and Xception have resulted in accuracy of only 77 percent and 93 percent respectively while DensNet201 and VGG19 performed decently having the same accuracy of 95 percent. The method discussed in [10] closely performed as compared to our model but failed in terms of recall and F1-score. Moreover it needs higher supports (1160) than other methods as shown in the Table 6.

Table 6: Accuracy of different methods in classifying spam and ham images.

Model	**Precision**	**Recall**	**F1-Score**	**Support**
VGG19 [17]	0.95	0.95	0.95	1154
CNN [10]	0.96	0.95	0.95	1160
Xception [18]	0.93	0.93	0.93	1154
ResNet251 [26]	0.78	0.77	0.77	1154
DenseNet201 [27]	0.95	0.95	0.95	1154
Our Model	**0.96**	**0.96**	**0.96**	**1154**

5 Conclusion

In this work the effectiveness of the proposal 3-stream neural network in classifying whether a given image is spam or ham is studied. Individual effectiveness of the features like SIFT and HOG is studied and appropriately fused in our 3-stream model for improving the accuracy of transfer learning methods. We have employed the transfer learning technique by employing pre-trained CNN architecture VGG-19 in our model. We have experimented with the effectiveness of the SIFT feature vs SIFT image with our test image dataset. Similarly, we have tested the effectiveness of HOG feature vs HOG image in spam image classification. We have also employed the Bag-of-words model to capture all key-points of the SIFT features. Our model performs better than previous methods. It can be concluded that in order to build a better image spam classifier, additional feature information like SIFT and HOG should also be used in model training and testing. In future works, we will try to extract local features from images using faster RCNN and study its effectiveness.

References

1. "Radicati group email statistics report 2021-25": accessed:2 jun 2022.[online]: Available:https://www.radicati.com/wp/wp-content/uploads/2021/Email_Statistics_Report,_2021-2025_Executive_Summary.pdf

2. C.-C. Lai and M.-C. Tsai: “An empirical performance comparison of machine learning methods for spam e-mail categorization,” in Fourth International Conference on Hybrid Intelligent Systems (HIS’04). IEEE,2004, pp. 44–48.
3. Image spam classification using OCR technique. January 2017. Sunita V. Dhavale: Available: https://www.researchgate.net/publication/315388437_Image_ Spam_Filters_Based_on_Optical_Character_Recognition_OCR_Techniques
4. Amara Dinesh Kumar; Vinayakumar R; Soman KP. Deep Learning based Image Spam Detection 3 Oct 2018: Available:https://arxiv.org/abs/1810.03977
5. R. Vinayakumar, K. Soman, P. Poornachandran, and S. Akarsh, “Application of deep learning architectures for cyber security,” in Cybersecurity and Secure Information Systems. Springer, 2019, pp. 125–160.
6. S. Venkatraman, M. Alazab, and R. Vinayakumar, “A hybrid deep learning image-based analysis for effective malware detection,” Journal of Information Security and Applications, vol. 47, pp. 377–389, 2019.
7. R. Vinayakumar, M. Alazab, K. Soman, P. Poornachandran, and S. Venkatraman, “Robust intelligent malware detection using deep learning,” IEEE Access, vol. 7, pp. 46 717–46 738, 2019.
8. V. S. Mohan, R. Vinayakumar, K. Soman, and P. Poornachandran, “Spoof net: syntactic patterns for identification of ominous online factors,” in 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018, pp. 258–263.
9. R. Vinayakumar, M. Alazab, K. Soman, P. Poornachandran, A. AlNemrat, and S. Venkatraman, “Deep learning approach for intelligent intrusion detection system,” IEEE Access, vol. 7, pp. 41 525–41 550, 2019.
10. https://ieeexplore.ieee.org/document/9044249
11. https://doi.org/10.1007/s42979-021-00762-x
12. https://www.researchgate.net/publication/306186011_Facial_Expression _Recognition_Using_a textunderscore Hybrid_CNN-SIFT_Aggregator
13. https://www.researchgate.net/publication/355152280_Image-Based_Malware_ Classification_Using_VGG19_Network_and_Spatial_Convolutional_Attention
14. https://ieeexplore.ieee.org/document/7860934
15. https://link.springer.com/article/10.1007/s00779-018-1168-8
16. ADAM: Kingma D, Adam BJ (2014) A method for stochastic optimization[J]. Computer Science
17. Bansalhttps://doi.org/10.1007/s12652-021-03488-z
18. Xception: 7 Oct 2016: François Chollet: arXiv:1610.02357v3
19. https://www.researchgate.net/publication/346534797_Color_Model _Based_Convolutional_Neural_Network_for_Image_Spam_Classification
20. https://arxiv.org/abs/2204.01710
21. https://docs.opencv.org/4.x/da/d22/tutorial_py_canny.html
22. https://www.researchgate.net/publication/338511771_An_Overview _of_Bag_of_WordsImportance_Implementation_Applications_and_Challenges
23. Introduction to machine learning with applications in information security. Chapman & Hall/CRC
24. M. Dredze, R. Gevaryahu, and A. Elias-Bachrach, “Learning fast classifiers for image spam.” in CEAS, 2007, pp. 2007–487.
25. Y. Gao, M. Yang, X. Zhao, B. Pardo, Y. Wu, T. N. Pappas, and A. Choudhary, “Image spam hunter,” in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2008, pp. 1765–1768.
26. ResNet251: Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun:10 Dec 2015: arXiv:1512.03385
27. DenseNet201:https://arxiv.org/abs/1608.06993v5