Development of a Machine Learning Model for Image-based Email Spam Detection

*[1]Christopher U. Onova, and [2]Temidayo O. Omotehinwa
[1]Department of Mathematical Sciences, Achievers University, Owo, Nigeria
[2]Department of Mathematics and Computer Science, Federal University of Health Sciences, Otukpo, Nigeria
{cuonova|oluomotehinwa}@gmail.com

ORIGINAL RESEARCH ARTICLE
Received: 25-OCT-2021; Reviewed: 13-DEC-2021; Accepted: 20-DEC-2021
http://dx.doi.org/10.46792/fuoyejet.v6i4.718

Abstract - Combatting email spam has remained a very daunting task. Despite the over 99% accuracy in most non-image-based spam email detection, studies on image-based spam hardly attain such a high level of accuracy as new email spamming techniques that defeat existing spam filters emerges from time to time. The number of email spams sent out daily has remained a key factor in the continued use of spam. In this paper, a simple convolutional neural network model, 123DNet was developed and trained with 28,929 images drawn from 2 public datasets and a Personally Generated dataset. The model was optimized to the least set of layers to have 1 input layer, 2 embedded Convolutional layers as a hidden layer, and 3 neural network layers. The model was tested with a total of 4,339 images of the three dataset samples and then with a separate set of 1,200 images to test performance on never-seen-before images. A Classification Performance analysis was carried out using the confusion matrix. Performance metrics including Accuracy, Precision, True Negative Accuracy, Sensitivity, Specificity, and F1 Measure were computed to ascertain the model's performance. The Model returned an F1 Score of 97% on a public dataset's test sample and 88% on Never-seen-before test samples outperforming some pre-existing models while performing significantly well on the newly generated image test samples. It is recommended that a model that performed so well with new never-seen-before spam images be integrated into spam filtering systems.

Keywords- Convolutional Neural Network, Deep Learning, Image, Spam Detection

1 INTRODUCTION

The internet has changed the way we work and live through several created platforms. Email is one such platform. An email is an electronic framework by which messages are transmitted from one user to the other (Bhuiyan, Ashiquzzaman, Juthi, Biswas, & Ara, 2018). Active email users have exceeded 4 billion and existing functional email accounts are estimated to be over 7 billion (99Firms, 2021). Thus, it has become very difficult for communication to take place among business concerns today without the use of email.

The exponential growth and popularity of use, coupled with very high reachability, and a significantly low cost of operation, has made the email a more economical messaging platform for sending a new type of email called spam (Shandilya, Polash & Shiva 2014). Spam emails also called junk mails, are unsolicited bulk e-mails, sent to random recipients in large quantities, with commercial, fraudulent, or malicious intentions (Khawandi, Abdallah, & Ismail, 2019; Sharmin, Di Troia, Potika, & Stamp, 2020). Spammers (senders of spam emails) have evolved several spamming techniques to fool existing spam filters with identifiable weaknesses. This includes manipulating plain texts by deliberate misspelling, obfuscating texts through the use of transform tools to change the look of text characters, and more recently, embedding texts (that are obfuscated, misspelt or not) in an image just to ensure that spam filters always record as much misclassification of spam as ham and vice versa as possible. Spamming sophistication improved to the point where spammers began to use randomization algorithms to generate images into which messages are embedded.

Although there are ongoing efforts at combating the spam menace, Guzella and Caminhas (2009), in a review, reported that some countries put in place, legislation against spam emails, but enforcement were weakened because spam emails are not usually sent from the geographic locations of these countries hence, making the process of tracking the actual senders of spam emails and by extension effective enforcement of these legislations, a very difficult task. This resulted in the need to take the battle to the cyberspace where spams originate, giving rise to the research efforts at developing various spam filters including those for image-based spams to identify legitimate emails and accurately detect spams, through analyzing incoming electronic mail.

As image-based spam got more complex, the Optical Character Recognition (OCR) became a very ineffective algorithm. Researchers eventually evolved a combination of several algorithms to create algorithms such as Deep Learning (DL) Algorithm which is an advancement of the Neural Network that is currently being deployed alongside other algorithms to tackle very complex trends in image-based spams. DL algorithms are a subset of Machine Learning (ML) that mimics the neural

*Corresponding Author

Section B- ELECTRICAL/ COMPUTER ENGINEERING & RELATED SCIENCES
Can be cited as:
Onova C.U., and Omotehinwa T.O. (2021): Development of a Machine Learning Model for Image-based Email Spam Detection, FUOYE Journal of Engineering and Technology (FUOYEJET), 6(4), 336-340. http://dx.doi.org/10.46792/fuoyejet.v6i4.718

functionality of the human brain. The DL approach was designed to mitigate the weaknesses of other ML algorithms.

The most prominent weakness of other ML algorithms is the popular "Curse of Dimensionality" where the algorithm becomes less effective as the number of features it has to analyse becomes very large. Although DL provides a better option when it comes to working with data with very complex features, it usually requires learning with an extremely large amount of dataset to attain accuracy levels compared to those of its predecessor ML algorithms.

The Convolutional Neural Network (CNN) also referred to as the Deep Learning model has proven to be effective in image classification problems. This is essential because images are complex data to manage when it comes to classification owning to the multifaceted nature of features required for perfect or near-perfect prediction.

2 Related Work

Singh (2018) in his work explored and evaluated four deep learning techniques that detect image spams, they include neural networks, deep neural networks, convolution neural networks, and transfer learning. The Dredze Dataset, Image Spam Hunter (ISH) dataset, an "Improved" dataset as well as the combination of these datasets formed the four categories of the datasets used in this work, to explore the robustness of the proposed model and to ascertain how well it performed bearing in mind the challenges created by spammers to outsmart image spam detection. The experimental results were compared with an existing VGG19 transfer learning model for detecting image spams. In the study, the results of the accuracy analysis for the dataset were 98.78% for ISH, 98.95% for the Dredze dataset, and 96.82% accuracy was recorded for the combination of the Dredze and Spam archive datasets; with 95.63% record as the accuracy for the combination of all datasets.

Yang, Liu, Zhou, and Luo (2019) proposed what they called the Multi-Modal Architecture based on Model Fusion (MMA-MF). This model utilized Long Short-Term Memory (LSTM) to filter spam. In this study, the LSTM and CNN models were used to process text and image components of emails separately to achieve two classification probability values that were incorporated into a fusion model to determine if the email was spam or ham. Also, a grid search optimization method was used to get the most appropriate hyperparameters for the MMA-MF. Performance evaluation of the model was carried out using a k-fold cross-validation method. The result of this study shows that the MMA-MF model achieved accuracies ranging between 92.64% and 98.48%.

Kim, Abuadbba, and Kim (2020), in a study, proposed DeepCapture a CNN-XGBoost framework comprising of eight layers with large training samples to show the feasibility of addressing the issue of performance degradation against entirely new image spam email. They evaluated performance with a dataset consisting of 6,000 spam and 2,313 non-spam image samples. Result achieved an F1-score of 88%, which was a 6% improvement over the best existing spam model CNN-SVM that at the time recorded an 82% F1-score.

Mohammad (2020), in a study, presented an enhanced model for ensuring a lifelong spam classification model called the Ensemble-based lifelong Classification using Adjustable Dataset Partitioning (ELCADP). The model is designed to handle what was perceived and referred to as "catastrophic forgetting" which is a scenario whereby new spam detection systems were unable to detect a recycled spamming approach because it was designed with no recourse to the body of knowledge acquired from previous spam attacks. Apart from the phenomenon of "catastrophic forgetting", the study focused on "concept drift" as a means of detecting a possible introduction of a change in the spamming methods. The model was able to evolve a new set of rules in response to any new approach introduced by a spammer based on the differences detected. For evaluation purposes, the overall performance of the suggested model is contrasted against various other stream mining classification techniques. The results proved the success of the suggested model as a lifelong spam email classification method. The identified gap in this authors work is that it did not capture how effective the model is when it comes to image-based spam, however, it is pertinent to bear in mind the two phenomena raised - the issue of "concept drift" and "catastrophic forgetting" so a model developed should be capable of detecting both and attempt to recycle an old spamming approach or detect the evolution of a new spamming approach.

Sharmin, Di Troia, Potika, & Stamp, (2020) studied the problem of image spam detection, based on image analysis, where they applied convolutional neural networks (CNN). In the study, comparisons were made between results of other machine learning techniques and that of the study, and the results of previous related work. Two categories of datasets were considered, included real-world image spam datasets and challenging image spam-like datasets created specifically for the study. Results showed an improvement on previous work as a new feature set consisting of a combination of the raw image and canny edges, were used. Srinivasan, Vinayakumar, Vishvanathan, Krichen, Noureddine, Anivilla, and Soman (2020) proposed a hybrid deep convolution neural network (DCNN) consisting of two convolution neural networks CNN1 and CNN2. In the study, the DCNN was trained with three datasets while also exploring transfer learning through engaging preexisting frameworks such as VGGC19, ImageNet in the training phase. They presented a result that showed an F1-score of 97.4% and an accuracy of 97.1%

From the studies reviewed, it is evident that active research work has been ongoing to address the issue that focuses on spam detection with more successes recorded in text-based spam. Various machine learning models have been deployed to combat email spam and specifically image-based spam. The weakness of most of these models stems from the fact that they are feature-based which in itself, is a requirement for objectively

http://dx.doi.org/10.46792/fuoyejet.v6i4.718 http://journal.engineering.fuoye.edu.ng/

building any model. However, because the characteristics of the features are known and are easily determined, spammers from time to time, come up with methods that seem to modify any feature that the models have gained mastery to effectively detect spam.

With the multiplicity of features came the need for models that could handle the related complexities. Most machine learning models need to be enhanced to accommodate large datasets with multiple complex features, as well as overcome its "curse of dimensionality". The advent of the neural networks was a breakthrough, with CNN being the most effective for working with images. Kim, Abuadbba, and Kim (2020) had reported that CNN and SVM had recorded a better F1-score before their study, lending credence to the fact that CNN is a model of choice as far as image-based spam is concerned. The strength of CNN is two folds - the ability to extract features peculiar to each image sample no matter how many they are, and the fact that, given the right hardware specification, the model only gets better with increasing dataset samples size.

Having generally identified spam detection as a drift problem (Dada et. al. 2019), and a model's capacity to handle new spams while not forgetting old spam methods (Mohammed, 2020), the identified gap that has necessitated this research was an attempt to explore how effective a simple CNN model could be at detecting new image-based spam email. Is it possible to outperform the existing models with a simple convolutional neural network (CNN) and yet be able to properly detect new spam emails? This study shall attempt to answer this question.

3 METHODOLOGY/ EXPERIMENTAL SETTINGS

The Google Colaboration research platform and Google Drive served as model implementation interface and dataset storage respectively, in this study.

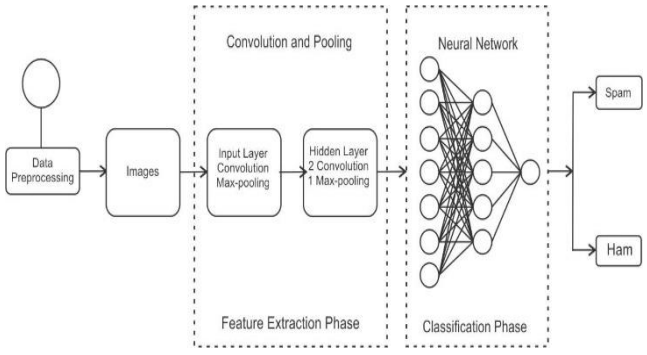

Fig. 1: The architecture of the Model

Figure 1 and Figure 2 are primarily about the model. There is the Feature extraction phase and the Classification phase. The first phase of the model comprised of an input layer having a 16 kernel, convolutional layer, a Rectified Linear Unit (ReLU) activation function, and a 2x2 max-pooling, next is a hidden layer having two convolution layers with 32 and 64 kernels, each having a ReLU activation function and this second layer terminating with one 2x2 max pooling. The flattening layer precedes the classification phase. It supplies the neural network (which is the centre of the classification phase) with a single vector stream of inputs. The Neural Network is divided into the three layers having 128 and 64 nodes respectively, and then a single node output having the sigmoid activation function to determine which class the image inputted into the model belongs to – whether it is 'spam' or 'not-spam'.

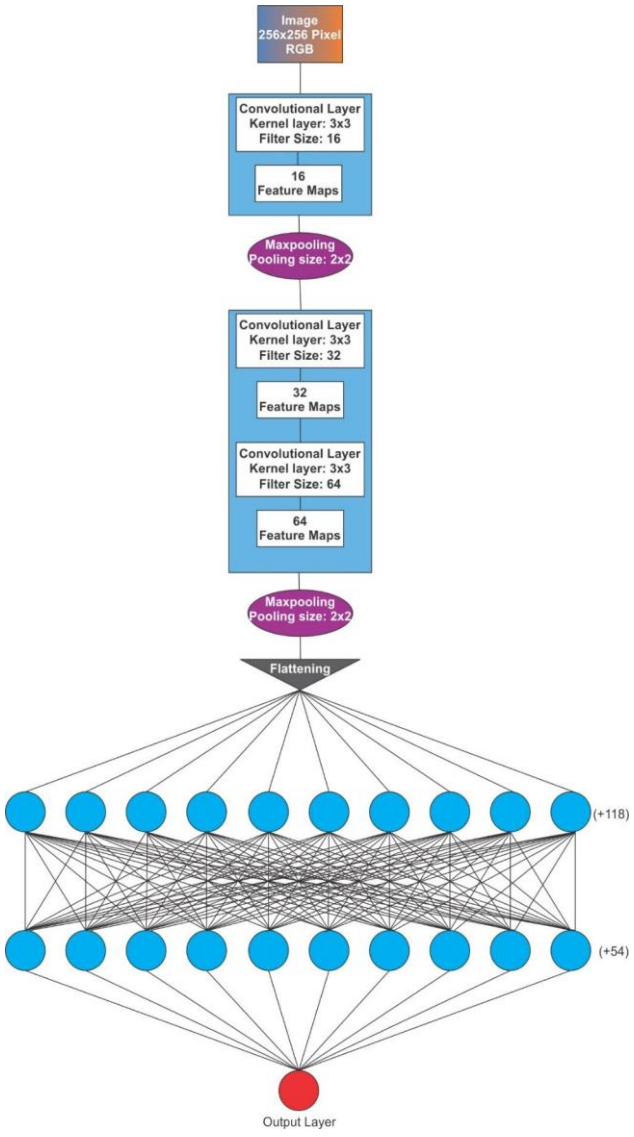

Fig. 2: Schematic diagram of 123DNet Model.

The datasets include; two public datasets – the Dredze image dataset (20,376 image files) (Dredze, Gevaryahu, and Elias-Barach, 2007); Image SpamHunter (ISH) dataset (1,740 image files) (Gao and Zhao 2009); and a personally generated dataset PERS-G (6,813 image files extracted from private Gmail accounts and Organisational webmail accounts). A separate set of 1200 images files were extracted solely to test the model's predictive capacity for never-seen-before images.

The image files utilised in this study underwent pre-processing to cater for Data Quality Assurance, Feature encoding, and Data split. Non-image file extensions were dropped and all the files of the datasets were converted to the jpeg format for uniformity as ISH, one of the datasets has only one file type (jpeg). With data split each dataset

http://dx.doi.org/10.46792/fuoyejet.v6i4.718 http://journal.engineering.fuoye.edu.ng/

was split into 70:15:15; for training, cross-validation and testing, respectively.

Table 1. Model Parameters

Parameter	Value
Number of Epoch	30
Batch Size	10
Step per Epoch	Depends on Batch size and Total number of Dataset Image
Learning Rate	0.001 – 0.0001
Drop out	30% to 50%

Parameters such as learning rate, step per epochs, and drop out were varied within the range stated in Table 1.0 during experimental phases to improve the model and balance 'over-fitting' before stable values were arrived at for the training of the model with all three datasets.

3.1 Performance Evaluation Metrics

Using the confusion matrix, the following classification metrics were computed. These are:

i. Accuracy
ii. Positive Prediction Value (PPV) or Precision,
iii. True Negative Accuracy,
iv. Specificity or True Negative Rate (TNR),
v. Sensitivity or True Positive Rate (TPR)
vi. F1 Score

Table 2 The Confusion Matrix Model

	p	*n*
Spam	***TP***	***FP***
Ham	***FN***	***TN***

Where *p = Spam (the positive sample),*
n = Ham (the negative sample),
i = Instance and P = Prediction,
Such that;
For i = p and P = p; is True Positive (TP)
i = p and P = n ; is False Negative (FN)
i = n and P = p; is False Positive (FP)
i = n and P = n; is True Negative (TN)

$$p = TP + FN \quad (1)$$

$$n = FP + TN \quad (2)$$

$$\text{Accuracy} = \frac{TP+TN}{p+n} \quad (3)$$

$$\text{Precision (PPV)} = \frac{TP}{FP+TP} \quad (4)$$

$$\text{True Negative Accuracy} = \frac{TN}{FN+TN} \quad (5)$$

$$\textit{Specificity or True Negative Rate (TNR)} = \frac{TN}{n} \quad (6)$$

$$\text{Sensitivity or True Positive Rate (TPR)} = \frac{TP}{p} \quad (7)$$

F1 Score derived from Equations 4 and 7 as

$$\text{F1 Score (F-Measure)} = \frac{2(PPV)xTPR}{PPV+TPR} \quad (8)$$

4 Results and Discussion

The model was trained with the three datasets and tested in two concurrent phases using 15% of the dataset for spam and 'not spam' files for each of the datasets as test samples on one hand, and a 1,200 image comprising a balanced class test dataset with 600 ham, and 600 not spam images, on the other.

Table 3. The Confusion matrix for Model Test with Dredze Dataset's Test sample

		(Predicted Class) *p*	*n*
(Actual Class)	*Spam*	*2,646*	*37*
	Ham	*107*	*266*

Table 3 shows the confusion matrix computation of test results for model's testing with the Dredze dataset's test sample. Table 4 and Table 5 show the result for ISH and PERS-G datasets respectively in their confusion matrix.

Table 4. The Confusion matrix for Model Test with ISH Dataset's Test sample

		(Predicted Class) *p*	*n*
(Actual Class)	*Spam*	*138*	*24*
	Ham	*2*	*97*

Table 5. The Confusion matrix for Model Test with PERS-G Dataset's Test sample

		(Predicted Class) *p*	*n*
(Actual Class)	*Spam*	*720*	*112*
	Ham	*14*	*175*

Table 6 shows the summary of results with values for performance metrics for all datasets. The F1 Score for the Dredze dataset stands as 97%.

Table 6. Summary of test results for dataset test samples

Performance Metric	*DataSets*		
	Dredze	*ISH*	*PERS-G*
Accuracy	*95%*	*90%*	*88%*
Precision	*99%*	*85%*	*86%*
True Negative Accuracy	*71%*	*98%*	*93%*
Specificity	*88%*	*80%*	*61%*
Sensitivity	*96%*	*99%*	*98%*
F1 Score	*97%*	*91%*	*92%*

Table 7. The Confusion matrix for Model Test with 1200 image samples after training with Dredze Dataset

		(Predicted Class) *p*	*n*
(Actual Class)	*Spam*	*583*	*185*
	Ham	*17*	*415*

http://dx.doi.org/10.46792/fuoyejet.v6i4.718 http://journal.engineering.fuoye.edu.ng/

Table 8. The Confusion matrix for Model Test with 1200 image samples after training with ISH Dataset

		(Predicted Class)	
		p	*n*
(Actual Class)	*Spam*	*591*	*145*
	Ham	*9*	*455*

Table 9. The Confusion matrix for Model Test with 1200 image samples after training with PERS-G Dataset

		(Predicted Class)	
		p	*n*
(Actual Class)	*Spam*	*572*	*142*
	Ham	*28*	*458*

Table 7 shows the confusion matrix computation of the test result for the model's testing with the 1200 image sample after the model's training with the Dredze dataset. Tables 8 and 9 show the result in a confusion matrix for model's testing with the 1200 image samples after training the model with ISH and PERS-G dataset respectively.

Table 10. Summary of test results outcome for the 1200 separate images

Performance Metric	*Datasets*		
	Dredze	*ISH*	*PERS-G*
Accuracy	*83%*	*86%*	*86%*
Precision	*76%*	*79%*	*80%*
True Negative Accuracy	*96%*	*98%*	*94%*
Specificity	*69%*	*75%*	*76%*
Sensitivity	*97%*	*99%*	*95%*
F1 Score	*85%*	*88%*	*87%*

Table 11. Class Distribution and F1-Score for test on dataset's test samples

Dataset	Class Distribution		% of Dataset in the total volume of training data	F1-Score
	Spam	Ham		
Dredze	90.1%	9.9%	70.43	97%
ISH	53.34%	46.66%	6.01	91%
PERS-G	71.8%	28.2%	23.55	92%

Table 12. Class Distribution and F1-Score for test on Never-Seen-Before test sample

Dataset	Class Distribution		% of Dataset in the total volume of training data	F1-Score
	Spam	Ham		
Dredze	90.1%	9.9%	70.43	85%
ISH	53.34%	46.66%	6.01	88%
PERS-G	71.8%	28.2%	23.55	87%

There is a correlation between the result of the model's performance on the never-seen-before image and the class distribution. Of the three datasets, the ISH has the least marginal difference in its class distribution, next is the PERS-G dataset. The 1200 image test dataset is a class balanced sample whose test result shows that the model spam predictive capacity is strong even with an imbalance training sample like the Dredze dataset. This implies that the model can respond to new image-based spam threats.

5 CONCLUSION

The model was able to achieve an 88% F1-score on the never-seen-before images and 97% F1-score on the dataset's test samples, thus performing at per with some and outperforming other models. The study has shown that with improved parameter tuning, image augmentation, and a marginal increase in computational power, a simple model could be enhanced to combat new Spam threats. A convolutional neural network model as 123DNet should be integrated into spam filtering systems. It is recommended that this model be reviewed and exposed to larger datasets to determine if its performance could be further improved.

REFERENCES

99Firms (2021). *How many email users are there?* Retrieved from https://99firms.com/blog/how-many-email-users-are-there/#gref

Bhuiyan, H., Ashiquzzaman, A., Juthi, T., Biswas, S., & Ara, J. (2018). *A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques*. Global Journal of Computer Science and Technology

Dada, E.G., Bassi, J.S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). *Machine learning for email spam filtering: Review, approaches, and open research problems*. Heliyon, 5(6), e01802. doi:10.1016/j.heliyon.2019.e01802

Dredze, M., Gevaryahu, R. & Elias-Bachrach, A. (2007). *Learning fast classifiers for image spam*. Retrieved from https://www.cs.jhu.edu/~mdredze/publications/image_spam_ceas07.pdf

Gao, Y., Yang, M., & Zhao, X. (2009), *Image spam hunter*. Retrieved from https://users.cs.northwestern.edu/~yga751/ML/ISpamHunter.pdf

Guzella, T. S. & Caminhas, W. M. (2009). *A review of machine learning approaches to spam filtering*. Expert Systems with Applications, 36 10206– 10222, Elsevier Ltd.

Khawandi, S., Abdallah, F. & Ismail, A. (2019). *A survey on image spam detection techniques*. Dhinaharan Nagamalai et al. (Eds) : COMIT, AISCA – 2019, 13–27, 2019. DOI: 10.5121/csit.2019.90102

Kim, B., Abuadbba, S., & Kim, H. (2020). *DeepCapture: Image spam detection using deep learning and data augmentation*. Retrieved from https://arxiv.org/abs/2006.08885

Mohammad, R. M. A. (2020). *A lifelong spam emails classification model*. Applied Computing and Informatics. A preprint. https://doi.org/10.1016/j.aci.2020.01.002

Sharmin, T., Di Troia, F., Potika, K., & Stamp, M. (2020). *Convolutional neural networks for image spam detection*. Information Security Journal: A Global Perspective, 1–15. doi:10.1080/19393555.2020.1722867

Singh, A.P. (2018). *Image spam classification using deep learning*. Master's Thesis and Graduate Research, SJSU ScholarWorks, San Jose State University.

Srinivasan, S., Vinayakumar, R., Vishvanathan, S., Krichen, M., Noureddine, D., Anivilla, S. & Soman, K.P. (2020). *Deep convolutional neural network-based image spam classification*. 6th Conference on Data Science and Machine Learning Applications *(CDMA)*, Riyadh, Saudi Arabia, 2020, pp. 112-117.

Yang, H., Liu, Q., Zhou, S., & Luo, Y. (2019). *A spam filtering method based on multi-modal fusion.* Applied Sciences, 9(6), 1152. doi:10.3390/app9061152