Research Article BEU Fen Bilimleri Dergisi 15(2): 557-565 (2026) DOI: 10.17798/bitlisfen.1779957

BİTLİS EREN UNIVERSITY
FEN BİLİMLERİ DERGİSİ
ISSN: 2147-3129 / e-ISSN: 2147-3188

ViTEmIS: VISION TRANSFORMER BASED EMAIL IMAGE SPAM DETECTOR

Sultan ZAVRAK

Duzce University, Computer Engineering Department, Düzce, Türkiye, sultanzavrak@duzce.edu.tr

Article Info

Received: *September 8, 2025*
Revised: *March 28, 2026*
Accepted: *March 31, 2026*

Keywords
Vision Transformer,
Image-based Email Spam Detection,
Spam Detection,
Email Spam.

ABSTRACT

Email spam is still a big threat to cybersecurity, and image-based spam makes it harder for traditional text-based spam detection methods to work. As impostors change their attack techniques, the need for creative ways to solve this problem is growing. This work presents ViTEmIS, a ViT-Base fine-tuning–based approach for detecting image-based email spam, together with a systematic analysis of duplicate-image effects on performance reporting across benchmark datasets. In addition, the proposed model is compared with the methods proposed in the previous studies. According to evaluation results, the proposed model shows promising results in terms of accuracy, precision, recall, and F1 on the different datasets. Across four benchmark datasets, ViTEmIS achieved 94.00–99.42% accuracy with duplicates included and 90.85–99.42% after duplicate removal, with corresponding F1 scores of 96.47–99.46% and 93.70–99.44%, respectively. The contributions are: (i) an evaluation of a ViT-Base fine-tuning approach for image-based email spam detection on multiple benchmarks, (ii) a systematic analysis of how duplicate images affect reported performance, and (iii) a comparative discussion with prior work with explicit modality caveats.

1. INTRODUCTION

E-mail has become an essential communication medium in the modern digital era. But it has emerged as a target for malicious actors who spread unwanted messages to innocent recipients. Numerous detection algorithms have been developed to address this issue. Usually, most of these algorithms employ natural language processing (NLP) techniques [1], [2] to analyze the textual content of emails and try to identify spam messages. Because spam emails sometimes contain graphics that evade text-based filtering, researchers have presented detection approaches [3], [4] that get the visual elements of email content as input.

Many methods have been investigated for image-based spam detection. Prior studies can be broadly grouped into: (i) approaches based on engineered visual features and classical classifiers such as support vector machines (SVMs) and eigenspace-based methods [5], [6]; (ii) robust visual modeling with ensemble methods (e.g., random forests with multigranularity features) [7]; and (iii) deep learning approaches using convolutional neural networks (CNNs) and transfer learning with pre-trained backbones [8], [9]. Some studies further integrate visual and textual cues (e.g., visual+text features) [10], including text derived via optical character recognition (OCR) combined with CNN features [11], and others incorporate additional signals such as links alongside image content [12]. Researchers have also explored system-level solutions that combine server-side filtering with client-side active learning [3].

Despite progress with CNN-based and hybrid OCR pipelines, two practical gaps remain: (i) reported performance can be inflated by duplicate or near-duplicate images that leak across training and testing splits, and (ii) evidence is limited on how transformer-based vision backbones perform on image-spam benchmarks under a duplicate-controlled protocol. This study addresses these gaps by evaluating ViT-Base fine-tuning across multiple benchmarks under both duplicate-included and duplicate-removed settings.

License: CC BY-NC 4.0

This work presents ViTEmIS, a ViT-Base fine-tuning–based approach for detecting image-based email spam, together with a systematic analysis of duplicate-image effects on performance reporting across benchmark datasets. This research provides three significant contributions: i) A comprehensive evaluation of the effectiveness and potential advantages of the ViT model architecture over traditional CNN methods is provided, which is carried out using multiple email image benchmark datasets. ii) The influence of duplicate entries on model performance is investigated. iii) A comparison with existing state-of-the-art techniques is presented by highlighting the reasonable performance of ViTEmIS. To support email security against evolving spam strategies, this study investigates ViT-based image spam detection and reports performance on established benchmark datasets under duplicate-controlled evaluation.

This paper is organized as follows: An overview of the ViT architecture and the datasets employed in the evaluation is provided in Section 2. Section 3 explains the experimental approach, which covers data preparation, model training, and assessment criteria. Subsequently, the experimental results are presented, and a comparative analysis with existing methods is conducted. Section 4 summarizes the findings, discusses the study's shortcomings, and suggests future research.

2. MATERIAL AND METHOD

2.1. Vision Transformer

The ViT architecture [13], [14] applies a multi-head self-attention mechanism to transform nonsequential spatial data into a sequence of small and fixed-size patches (e.g., 64x64). This method extracts comprehensive contextual information from an input image. The multi-head self-attention mechanism enables the selective focus on a variety of regions within the image and the understanding of the interdependencies across different segments over extended spans. This methodology circumvents the constraints that are inherent in tensor-based operations, including translation invariance and locally constrained receptive fields, which are common in conventional CNNs. Figure 1 illustrates a high-level schematic of the ViT architecture.

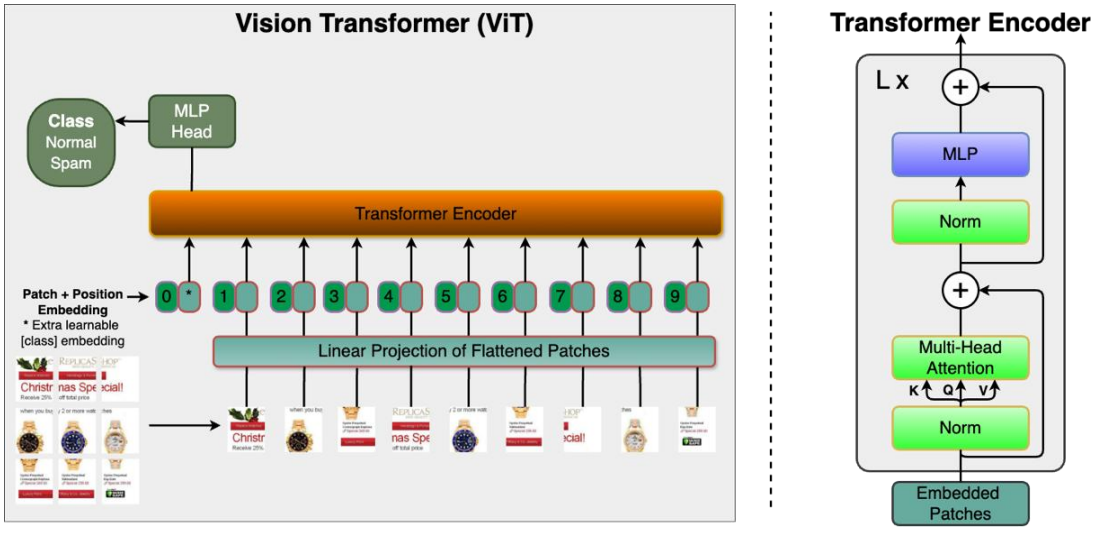


Figure 1. The overall architecture of ViT [13]

2.2. Dataset Description

For image-based email spam detection, it is essential to use diverse and realistic datasets for evaluating the proposed approach. Dredze et al. [4] created two datasets: one from personal spam and the other from public spam. The personal spam dataset was constructed by extracting all attached images (GIF, JPG, PNG, and BMP) from spam emails in 10 accounts and a catch-all filter over one month. Additional images were obtained from a subset of the SpamArchive corpus used by Fumera et al. [15]. Images from two user email accounts in the ham dataset, 2,359 images from the personal ham corpus, 2,173 images from the SpamArchive corpus, and 1,248 images from the personal spam dataset were combined into

DREDZE1 (personal ham and spam), DREDZE2 (personal ham and SpamArchive spam), and DREDZE3 (personal ham, spam, and SpamArchive spam). The personal spam archive comprised mostly images of stock market scams, and the SpamArchive corpus included a variety of images of advertisements.

Table 1. The number of images in the datasets (duplicates included)

	Dataset name	ham	spam	all
DREDZE1	Personal (ham+spam) [4]	1821	3213	5034
DREDZE2	Personal (ham)+SpamArchive (Spam) [4]	1821	12095	13916
DREDZE3	Personal (ham+spam)+SpamArchive (Spam) [4]	1821	15308	17129
ISH	ImageSpamHunter (ISH) [16]	810	929	1739

The ImageSpamHunter dataset created by Gao et al. [16] includes normal and spam images. Normal images, which are typical of social networking platforms, were randomly obtained from Flickr.com, whereas spam images were derived from actual spam emails.

Table 2. The number of images in the datasets (duplicates excluded)

	Dataset name	ham	spam	all
DREDZE1	Personal (ham+spam) [4]	1539	1262	2801
DREDZE2	Personal (ham)+SpamArchive (Spam) [4]	1537	4198	5735
DREDZE3	Personal (ham+spam)+SpamArchive (Spam) [4]	1537	5460	6997
ISH	ImageSpamHunter [16]	810	897	1707

Tables 1 and 2 summarize the distribution of the images with and without duplicates, respectively. The sample counts differ from those of the original datasets because of the corrupted images. Table 2 presents the revised datasets after systematic removal of duplicates. Duplicate removal was performed using an off-the-shelf duplicate-image detection utility that groups candidate duplicates based on visual content similarity rather than filenames. We used the utility's default similarity settings to generate candidate duplicate groups and then manually verified the candidates before removal. Only images confirmed as duplicates were removed. Corrupted/unreadable images were excluded prior to this process, which explains minor differences from the originally reported dataset sizes. In this study, "duplicates" refers to images that are exact or near-identical copies as flagged by the content-similarity utility and confirmed by manual inspection. The datasets varied significantly in terms of size and number of duplicate entries. DREDZE1 initially had 5,034 images, which were reduced to 2,801 after duplicate removal. DREDZE2 had 13,916 images, which were reduced to 5,735 after elimination. DREDZE3, the largest dataset, initially had over 17,000 images, which were reduced to less than 7,000 after excluding duplicates. ISH, the smallest dataset, contained approximately 1,700 images under both conditions, showing minimal duplicates.

To complement the quantitative dataset statistics in Table 2, Figure 2 presents representative ham and spam examples from the duplicate-removed DREDZE3 and ISH datasets. The figure is provided as a qualitative illustration of the visual characteristics of the benchmarks and does not affect the experimental protocol or the train/validation/test splits.



Figure 2. Representative ham and spam examples from duplicate-removed datasets (DREDZE3 and ISH).

3. EXPERIMENTAL RESULTS

3.1. Experimental Setup

The ViTEmIS model, which is based on the ViT-Base configuration [13], processes 224×224 pixel images with 16×16 patches. It includes 12 transformer layers, 768 hidden dimensions, 12 attention heads, and a multi-layer perceptron size of 3072, totaling 86 million parameters. ViTEmIS was developed by fine-tuning the ViT-Base model, pre-trained on ImageNet-21k, and fine-tuned on ImageNet 2012 [17], using the Hugging Face Transformers library. All images were converted to RGB when necessary and processed with a ViT-compatible pipeline. Inputs were resized for 224×224 ViT-Base patch16 inference. During training, lightweight augmentation was applied to improve generalization (random cropping); during evaluation, deterministic preprocessing (resize + center crop) was used. Inputs were normalized using the standard normalization associated with the ViT feature extractor in the Hugging Face Transformers implementation.

The experiments evaluate the impact of duplicate data on model performance by comparing the full datasets, which include duplicates, and refined datasets without duplicates. Four datasets (DREDZE1–3 and ISH) are evaluated under two conditions (duplicates included vs duplicates removed). For each dataset–condition pair, images from various directories were loaded, randomized for heterogeneity, and divided into training, validation, and test sets at a ratio of 70%, 10%, and 20%, respectively. The stratified splitting method was employed to maintain label distributions across subsets, ensuring a balanced representation throughout all phases and preserving class balance.

In order to enhance generalization and introduce variability, a few data augmentation techniques have been implemented on the training set to enhance model robustness, such as random cropping and normalization. To ensure consistency during evaluation, the images were resized and center-cropped to fit the input size of the model for the validation and test sets. Normalization was then performed.

A small validation-based sweep was performed around standard ViT fine-tuning settings (learning rate and weight decay), and the configuration providing the best validation performance and stable convergence was selected. The batch size was selected based on available GPU memory.

The model underwent fine-tuning with carefully selected hyperparameters listed in Table 3 and was enhanced through expert insights and extensive experimentation. The training process encompassed a maximum of 20 epochs, utilizing a batch size of 8 and a learning rate of 0.00008. A weight decay of 0.1 was implemented for regularization to mitigate overfitting. An early stopping procedure was employed

to stop training if the validation accuracy failed to improve by 0.01 over three successive epochs, thereby optimizing computational resources and minimizing overfitting.

Table 3. Training Parameters

Parameter	Value
Max epoch	20
Seed	137
Batch size	8
Learning rate	0.00008
Weight decay	0.1

3.2. Results and Discussion

The ability of ViTEmIS to detect email-image spam was evaluated using test data performance by employing four metrics: accuracy, precision, recall, and F1. Accuracy calculates correctly classified instances, recall measures the proportion of accurately identified spam emails, and precision calculates the fraction of correctly labeled spam among all spam classifications. The F1 score, which is the harmonic mean of the precision and recall, provides a comprehensive performance metric.

Table 4. The evaluation results of the ViTEmIS model (%) (duplicates included)

Dataset	Accuracy	Precision	Recall	F1
DREDZE1	96.52	97.64	96.88	97.26
DREDZE2	94.00	98.61	94.42	96.47
DREDZE3	95.79	97.40	97.90	97.65
ISH	99.42	98.93	100.0	99.46

ViTEmIS was assessed on datasets both with and without duplicates to examine the effect of duplicate entries. The experiments yielded significant insights into the impact of duplicates on the model's accuracy, precision, recall, and F1 score.

Table 5. The results of the ViTEmIS model (%) (duplicates excluded)

Dataset	Accuracy	Precision	Recall	F1
DREDZE1	97.33	97.98	96.04	97.00
DREDZE2	90.85	94.33	93.09	93.70
DREDZE3	92.64	96.08	94.41	95.24
ISH	99.42	99.44	99.44	99.44

Table 4 presents the results of the ViTEmIS model for the datasets that encompass duplicate images. The model showed remarkable performance across all the datasets. The model achieved an accuracy of 96.52% with precision, recall, and F1 scores all above 96% on the DREDZE1 dataset. The model had a slightly lower accuracy of 94.00%, but outstanding precision at 98.61% and an F1 score of 96.47% on the DREDZE2 dataset. This indicates that the model has strong spam identification ability. On the DREDZE3 dataset, the model achieved an accuracy of 95.79% and a recall of 97.90%. This indicates strong performance in identifying spam images in this benchmark setting. The model showed nearly flawless results on the ISH dataset, with an accuracy of 99.42% and an F1 of 99.46%. These results suggest that duplicates may inflate performance metrics because of the repeated exposure to reinforcing patterns during training.

Table 5 presents the performance results of the ViTEmIS model on the de-duplicated datasets. On the DREDZE1 dataset, the accuracy of the model slightly improved to 97.33%, whereas the recall decreased to 96.04%. This indicates that a more conservative model is less prone to overfitting. The model showed a significant drop in accuracy of 90.85%, with reduced precision (94.33%) and recall (93.09%) on the DREDZE2 dataset. On the DREDZE3 dataset, the accuracy of the model reached 92.64% with an F1 of

95.24%. On the ISH dataset, the model exhibited minimal performance changes, likely due to the low incidence of duplicates. It is inferred that the model's confidence in spam prediction was increased by the presence of duplicate images in the previous dataset, as the model's performance primarily decreases after the duplicates are excluded, as indicated by the results of various datasets.

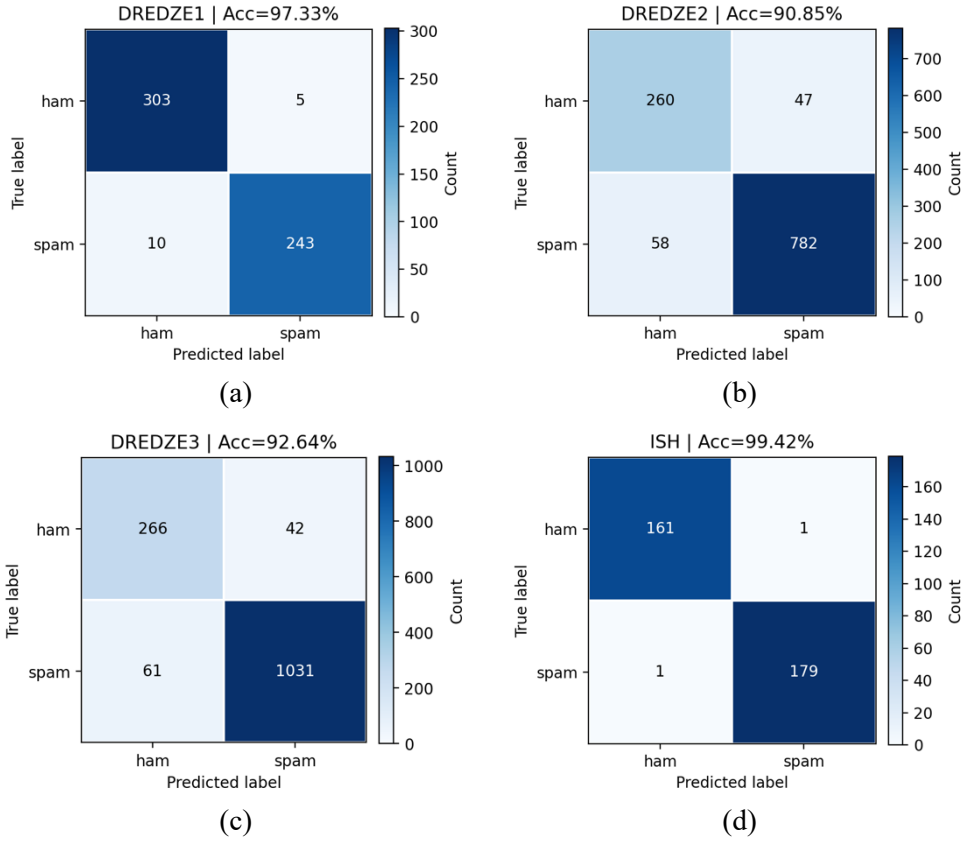


Figure 3. Confusion matrices for ViTEmIS on duplicate-removed test sets. (a) DREDZE1, (b) DREDZE2, (c) DREDZE3, and (d) ISH. Values denote the number of test samples in each category (TN, FP, FN, TP).

To analyze failure modes beyond aggregate metrics, Figure 3 reports confusion matrices for the duplicate-removed setting across all datasets.

Overall, the duplicate-removed setting remains highly accurate on DREDZE1 and ISH, with few misclassifications (DREDZE1: 5 false positives and 10 false negatives; ISH: 1 false positive and 1 false negative). In contrast, DREDZE2 exhibits the largest error counts (47 false positives and 58 false negatives), indicating greater overlap between ham and spam visual patterns and/or stronger domain variability in the underlying sources. DREDZE3 shows intermediate difficulty (42 false positives and 61 false negatives). These patterns suggest that performance degradation after duplicate removal is not uniform across datasets and is most pronounced when dataset composition is more heterogeneous, reinforcing the need for duplicate-controlled evaluation when reporting image-spam detection performance.

Removing duplicate entries from machine-learning training datasets is crucial for tasks such as email image spam detection, because it reveals the true generalization ability of the model. Although ViTEmIS performed quite well in both scenarios, training the model using de-duplicated datasets produced more reliable results.

Table 6 summarizes representative prior studies on image-based spam detection. Because prior work differs in input modality (e.g., some combine images with OCR-extracted text) and experimental protocol, Table 6 should be interpreted as contextual rather than as a strict ranking. Within the image-only setting on the evaluated benchmarks, ViTEmIS achieves competitive accuracy/F1, while hybrid

pipelines that incorporate OCR/text may report different performance under different assumptions and inputs.

Table 6. Comparison of ViTEmIS with Previous Studies.

Reference	Input modality	Main method/model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Dredze et al., 2007 [4] (duplicates excluded)	image-only	feature-based image spam classifier	DREDZE1	98	-	-	97
			DREDZE2	89	-	-	89
			DREDZE3	91	-	-	93
Shen et al. [7]	image-only	Random forest + multigranularity visual features	DREDZE1	96.8			
Annadatha et al., 2018 [6]	image-only	eigenspace/SVM-based + feature selection	ISH	97.25	-	-	-
Sharmin et al.,2020 [8]	image-only	CNN-based image spam detection	ISH	99.02	-	-	-
Makkar and Kumar, 2021 [12]	image + link features	deep learning combining link + image cues	ISH	96	-	-	-
Salama et al., 2023 [9]	image-only	transfer learning (ResNet50 best reported)	ISH	99.87	-	-	-
This study *(duplicates excluded)*	image-only	ViTEmIS	DREDZE1	97.33	97.98	96.04	97.00
			DREDZE2	90.85	94.33	93.09	93.70
			DREDZE3	92.64	96.08	94.41	95.24
			ISH	99.42	99.44	99.44	99.44

ViTEmIS achieved noteworthy results for the DREDZE datasets. Compared to the approach of Dredze et al. [4], ViTEmIS showed comparable or improved F1 results for the DREDZE1 and the DREDZE3 datasets. This indicates that the proposed model shows competitive performance despite higher computational costs. Additionally, ViTEmIS surpasses the accuracy of the method of Shen et al. [7] on the DREDZE1 dataset. This suggests ViTEmIS's global context capture is superior to the method that uses a random forest-based filter with multigranularity visual features.

On the ISH dataset, ViTEmIS demonstrated a near-perfect performance with 99.42% accuracy and a balanced F1 score of 99.44%, exceeding the eigenspace- and SVM-based techniques of Annadatha et al. [6]. Although Sharmin et al. [8] achieved similar accuracy using CNNs, ViTEmIS's balanced precision and recall were evidenced by the identical F1 score, indicating its potential for high and balanced performance. Despite Salama et al.'s [9] slightly higher accuracy with ResNet50 and transfer learning, this comprehensive evaluation, including duplicate analysis, offers greater insight. Additionally, Makkar and Kumar's [12] deep learning framework, which uses link and image content features, had lower accuracy than ViTEmIS on the ISH dataset. This suggests that learned features of the ViTEmIS model are more discriminative for image-based spam detection.

The ViTEmIS model shows competitive or superior performance compared to previous studies using various techniques, from handcrafted features and SVMs to CNNs. These results highlight the significant potential of the ViT architecture for this email image spam detection task. However, further research is required to explore the advantages of integrating textual features and thoroughly assessing generalizability across larger and more diverse datasets.

4. CONCLUSION

This study introduces ViTEmIS, a novel fine-tuned model that utilizes the ViT architecture for image-based email spam detection. The experiments with various benchmark datasets (DREDZE1, DREDZE2, DREDZE3, and ISH) show that the ViTEmIS model obtains high accuracy, precision, recall, and F1 scores, often exceeding those results of previous studies. The ViTEmIS model works well when there are no duplicates. When comparing datasets with and without duplicate samples on the same model,

the performance goes down in the datasets without duplicates. This shows how important it is to avoid duplicate samples while preparing datasets for email image spam detection.

This study is limited to established benchmark datasets derived from a small number of primary sources and does not include a dedicated evaluation on recent image spam. In addition, cross-dataset generalization experiments (training on one source and testing on another) are not reported. Finally, ViTEmIS is evaluated in an image-only setting without OCR/text fusion, and an exhaustive hyperparameter optimization is not conducted. These limitations motivate future work on more diverse and recent data, modality fusion, and cross-domain testing. To assess robustness and generalizability to evolving spam techniques, future work should evaluate ViTEmIS on larger, more diverse, and more recent datasets, including cross-dataset testing. Extensive grid searches for ViT configurations and hyperparameters may yield more robust and efficient models. By combining with text-based analysis, ViTEmIS may provide better spam-filtering by taking advantage of the text-based features. This work advances email security and sheds light on ViT architecture in email image spam detection.

Statement of Research and Publication Ethics

The study complies with research and publication ethics

Artificial Intelligence (AI) Contribution Statement

During the preparation of this study, the author used AI-assisted technologies to polish the article text. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

REFERENCES

[1] S. Zavrak and S. Yilmaz, “Email spam detection using hierarchical attention hybrid deep learning method,” Expert Syst. Appl., vol. 233, pp. 120977–120977, Dec. 2023, doi: 10.1016/J.ESWA.2023.120977.

[2] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, “A review of spam email detection: analysis of spammer strategies and the dataset shift problem,” Artif. Intell. Rev. 2022 562, vol. 56, no. 2, pp. 1145–1173, May 2022, doi: 10.1007/S10462-022-10195-4.

[3] Y. Gao, A. Choudhary, and G. Hua, “A Comprehensive Approach to Image Spam Detection: From Server to Client Solution,” IEEE Trans. Inf. Forensics Secur., vol. 5, no. 4, pp. 826–836, Dec. 2010, doi: 10.1109/TIFS.2010.2080267.

[4] M. Dredze, R. Gevaryahu, and A. Elias-Bachrach, “Learning Fast Classifiers for Image Spam,” presented at the International Conference on Email and Anti-Spam, 2007. Accessed: Aug. 18, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Learning-Fast-Classifiers-for-Image-Spam-Dredze-Gevaryahu/c6d975fcfd9a6f698e9eff0118fa3185dc8111e4

[5] B. Mehta, S. Nangia, M. Gupta, and W. Nejdl, “Detecting image spam using visual features and near duplicate detection,” in Proceedings of the 17th international conference on World Wide Web, in WWW ’08. New York, NY, USA: Association for Computing Machinery, Apr. 2008, pp. 497–506. doi: 10.1145/1367497.1367565.

[6] A. Annadatha and M. Stamp, “Image spam analysis and detection,” J. Comput. Virol. Hacking Tech., vol. 14, no. 1, pp. 39–52, Feb. 2018, doi: 10.1007/s11416-016-0287-x.

[7] J. Shen, R. H. Deng, Z. Cheng, L. Nie, and S. Yan, “On robust image spam filtering via comprehensive visual modeling,” Pattern Recognit., vol. 48, no. 10, pp. 3227–3238, Oct. 2015, doi: 10.1016/j.patcog.2015.02.027.

[8] T. Sharmin, F. Di Troia, K. Potika, and M. Stamp, “Convolutional neural networks for image spam detection,” Inf. Secur. J. Glob. Perspect., vol. 29, no. 3, pp. 103–117, May 2020, doi: 10.1080/19393555.2020.1722867.

[9] W. M. Salama, M. H. Aly, and Y. Abouelseoud, “Deep learning-based spam image filtering,” Alex. Eng. J., vol. 68, Jan. 2023, doi: 10.1016/j.aej.2023.01.048.

[10] T. Kumaresan, S. Saravanakumar, and R. Balamurugan, “Visual and textual features based email spam classification using S-Cuckoo search and hybrid kernel support vector machine,” Clust. Comput., vol. 22, no. 1, pp. 33–46, Jan. 2019, doi: 10.1007/s10586-017-1615-8.

[11] S.-G. Nam, Y. Jang, D.-G. Lee, and Y.-S. Seo, “Hybrid Features by Combining Visual and Text Information to Improve Spam Filtering Performance,” Electronics, vol. 11, no. 13, Art. no. 13, Jan. 2022, doi: 10.3390/electronics11132053.

[12] A. Makkar and N. Kumar, "PROTECTOR: An optimized deep learning-based framework for image spam detection and prevention," Future Gener. Comput. Syst., vol. 125, pp. 41–58, Dec. 2021, doi: 10.1016/j.future.2021.06.026.

[13] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: https://arxiv.org/abs/2010.11929v2

[14] E. Asadi Shamsabadi, C. Xu, A. S. Rao, T. Nguyen, T. Ngo, and D. Dias-da-Costa, "Vision transformer-based autonomous crack detection on asphalt and concrete surfaces," Autom. Constr., vol. 140, pp. 104316–104316, Aug. 2022, doi: 10.1016/J.AUTCON.2022.104316.

[15] G. Fumera, I. Pillai, and F. Roli, "Spam Filtering Based On The Analysis Of Text Information Embedded Into Images," J Mach Learn Res, Dec. 2006, Accessed: Aug. 18, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Spam-Filtering-Based-On-The-Analysis-Of-Text-Into-Fumera-Pillai/707cd6830fa9faffd6b8afb0a7f1b9a145e5c446

[16] Y. Gao et al., "Image spam hunter," in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 2008, pp. 1765–1768. doi: 10.1109/ICASSP.2008.4517972.

[17] "google/vit-base-patch16-224 · Hugging Face." Accessed: Nov. 21, 2024. [Online]. Available: https://huggingface.co/google/vit-base-patch16-224