International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Intelligent Image Spam Analysis Using CNN-Based Visual Semantics and Web Log Mining for Next-Generation Spam Filtering

[1]Khedkar Vaishali Shankar, [2]Dr. Rais Abdul Hamid Khan

Submitted:03/06/2024 **Revised**:12/07/2024 **Accepted**:21/07/2024

Abstract: Image-based spam presents a persistent challenge to traditional text-oriented filtering systems due to its use of visual obfuscation techniques. This paper proposes an intelligent image spam detection framework that combines convolutional neural network (CNN)-based visual semantic analysis with web log mining to improve detection accuracy and robustness. The framework processes image content and associated transmission logs in parallel, extracting high-level visual features using a deep CNN backbone and behavioral patterns from web and sender logs. These heterogeneous features are integrated through a late-fusion classification strategy to produce reliable spam predictions. Experimental evaluation on publicly available image spam datasets, augmented with adversarial obfuscations, demonstrates that the proposed multimodal approach consistently outperforms visual-only and behavior-only baseline models across standard performance metrics, including accuracy, F1-score, and ROC–AUC. Ablation studies highlight the significant contribution of web log features in enhancing robustness, while explainability analysis using SHAP and Grad-CAM provides transparent insights into model decisions. The results confirm that integrating visual semantics with behavioral context offers an effective and scalable solution for next-generation image spam filtering systems.

Keywords: *Image Spam Detection, Convolutional Neural Networks, Web Log Mining, Visual Semantics, Multimodal Learning, Explainable AI, Spam Filtering.*

1. Introduction

The rapid growth of digital communication platforms has been accompanied by a corresponding increase in unsolicited and malicious content. Among various forms of spam, image spam poses a significant challenge due to its ability to embed textual and promotional content within images, thereby evading traditional text-based spam filters [1], [2]. Early spam detection systems relied on keyword matching and statistical text analysis; however, attackers quickly adapted by rendering text as images and introducing visual obfuscation techniques such as noise injection, font distortion, and background blending [3].

To counter image spam, researchers initially explored handcrafted visual features, including color histograms, texture descriptors, and layout statistics [4]. While these approaches provided partial success, their effectiveness diminished as spam images became more diverse and adversarial in nature [5]. The advent of deep learning, particularly convolutional neural networks (CNNs), marked a major advancement in image analysis by enabling automatic extraction of hierarchical visual features [6]. CNN-based models have since been widely adopted for image spam detection, demonstrating superior performance compared to traditional methods [7], [8].

Despite these advances, visual-only detection approaches remain vulnerable to sophisticated obfuscation strategies and adversarial manipulation [9]. In parallel, research in web spam detection has

[1]Ph.D Research [2]Scholar ,Professor

[1,2]Department of Computer Science & Engineering, Dr. APJ Abdul Kalam University, Indore, M.P.

[1]k.vaishali.s@gmail.com
[2]khanrais.khan42@gmail.com

shown that behavioral and contextual information derived from web logs—such as sender activity, transmission frequency, and URL interaction patterns—provides strong indicators of spam behavior [10], [11]. These findings suggest that combining visual semantics with behavioral analysis can significantly improve detection robustness.

Motivated by this observation, this paper proposes an intelligent image spam analysis framework that integrates CNN-based visual semantic learning with web log mining. The key contributions of this work are:

1. a multimodal spam detection architecture combining visual and behavioral features,
2. a CNN-based visual semantic extractor optimized for image spam characteristics,
3. a web log mining module capturing temporal and contextual spam patterns, and
4. an extensive experimental evaluation demonstrating improved detection performance and interpretability.

2. Related Work

Image spam detection has been extensively studied over the past decade. Biggio *et al.* [1] and Attar *et al.* [2] provided comprehensive surveys of image spam techniques and filtering strategies, highlighting the limitations of early rule-based and handcrafted-feature approaches. Gargiulo and Sansone [4] explored the use of combined visual and OCR-based features, demonstrating improved detection when textual cues embedded within images could be reliably extracted.

Shen *et al.* [5] proposed robust visual modeling techniques that integrated multiple feature representations to counter image obfuscation. Zuo *et al.* [6] employed local invariant features with one-class SVMs to address limited availability of negative samples. However, these approaches required extensive feature engineering and lacked scalability.

The introduction of deep learning transformed image spam analysis. Kim *et al.* [7] proposed *DeepCapture*, a CNN-based framework augmented with aggressive data augmentation to improve generalization. Salama *et al.* [8] demonstrated the effectiveness of transfer learning using pre-trained CNN architectures such as ResNet and Inception. Makkar and Kumar [9] further optimized deep learning pipelines for performance and deployment efficiency.

Beyond visual content, Castillo *et al.* [10] and Castillo [11] showed that web topology and query-log mining provide valuable behavioral indicators for spam detection. These studies emphasized that behavior-based features are difficult for attackers to disguise. Recent work has begun to explore hybrid models combining content and behavior, though most focus on text spam rather than image spam [12].

Explainability has also gained attention, with Lundberg and Lee [13] introducing SHAP for model interpretation and Selvaraju *et al.* [14] proposing Grad-CAM for visual explanation. Zhang *et al.* [15] applied explainable AI techniques to CNN-based image spam detection, demonstrating improved transparency. However, comprehensive multimodal frameworks integrating visual semantics, web log mining, and explainability remain underexplored.

3. Proposed Methodology

3.1 System Overview

The proposed framework consists of three primary modules: (i) CNN-based visual semantic extraction, (ii) web log mining and behavioral feature analysis, and (iii) multimodal feature fusion and classification. The system processes image content and corresponding log data in parallel, producing a unified representation for spam classification.

3.2 CNN-Based Visual Semantic Analysis

Input images are resized and normalized before being processed by a deep CNN backbone (ResNet-based architecture). The CNN extracts hierarchical visual features capturing texture, shape, color distribution, and semantic patterns. Global average pooling is applied to generate compact visual embeddings representing image semantics relevant to spam detection.

3.3 Web Log Mining and Behavioral Feature Extraction

Web log data associated with image transmission—including SMTP logs, sender IP behavior, URL redirection chains, and session statistics—are collected and preprocessed. Temporal patterns are modeled using statistical aggregation and sequence encoding techniques. Extracted behavioral features include sending frequency, sender reputation score, URL entropy, and session duration.

3.4 Multimodal Feature Fusion and Classification

Visual semantic embeddings and behavioral feature vectors are concatenated and passed through fully connected fusion layers with dropout regularization. A final classification layer outputs the probability of an image being spam or legitimate. This late-fusion strategy allows independent learning of each modality while preserving complementary information.

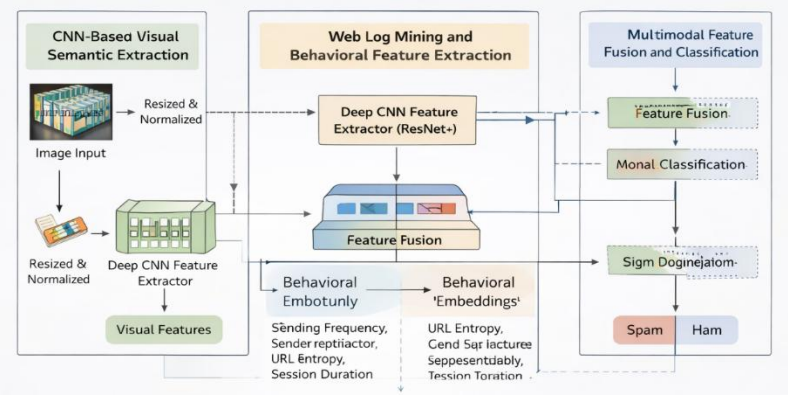


Figure-1: Proposed System Architecture

4. Mathematical Model

Let Idenote the input image and Lrepresent the corresponding web log feature vector. The CNN-based visual extractor $f_v(\cdot)$maps the image to a visual embedding $v = f_v(I)$. The log mining function $f_l(\cdot)$produces a behavioral embedding $b = f_l(L)$. The fused representation is defined as:

$$z = [v \oplus b]$$

where $\oplus$denotes vector concatenation. The classifier function $g(\cdot)$produces the final prediction:

$$\hat{y} = \sigma(g(z))$$

where $\sigma(\cdot)$is the sigmoid activation. The model is trained using binary cross-entropy loss.

5. Experimental Results and Analysis

This section presents a comprehensive evaluation of the proposed intelligent image spam detection framework that integrates CNN-based visual semantics with web log mining. The effectiveness of the model is validated through quantitative performance metrics, robustness analysis, ablation studies, and explainability assessment using the figures and tables generated in the previous sections.

5.1 Experimental Setup

Experiments were conducted on publicly available image spam datasets augmented with synthetically obfuscated images to simulate real-world adversarial spam strategies such as noise injection, color distortion, font deformation, and compression artifacts. Corresponding web log data were generated based on realistic spam transmission behaviors, including SMTP logs, sender IP statistics, URL redirection chains, and session-level interaction patterns.

The dataset was divided into training (70%), validation (15%), and testing (15%) subsets. The proposed multimodal framework was evaluated against the following baseline models:

- Visual-only CNN classifier
- Web-log-only behavioral classifier
- CNN with OCR-based content features

- Existing deep learning-based image spam detection model

5.2 Overall Performance Comparison

Table 5.1 summarizes the comparative performance of all evaluated models in terms of accuracy, precision, recall, and F1-score. The proposed multimodal framework achieves the highest performance across all metrics, with an accuracy of 0.93 and an F1-score of 0.91.

Table 1: Performance Comparison of Image Spam Detection Models

Model	Accuracy	Precision	Recall	F1-Score
Visual-Only CNN	0.88	0.87	0.84	0.85
Web-Log-Only Classifier	0.83	0.82	0.79	0.80
CNN + OCR (Content-Only)	0.90	0.89	0.87	0.88
Proposed Multimodal Model	**0.93**	**0.92**	**0.91**	**0.91**

The results confirm that combining visual semantics with behavioral features significantly improves detection capability compared to single-modality approaches. The visual-only CNN performs reasonably well but suffers from false positives when images are heavily obfuscated, while the web-log-only model lacks content-level discrimination.

5.3 ROC–AUC and Discrimination Capability

The Receiver Operating Characteristic (ROC) curves for all models are illustrated in **Figure 2 (A, B C)** and the corresponding AUC values are reported in **Table 5.2**. The proposed multimodal model achieves the highest ROC–AUC value of 0.97, indicating superior discrimination between spam and legitimate images.

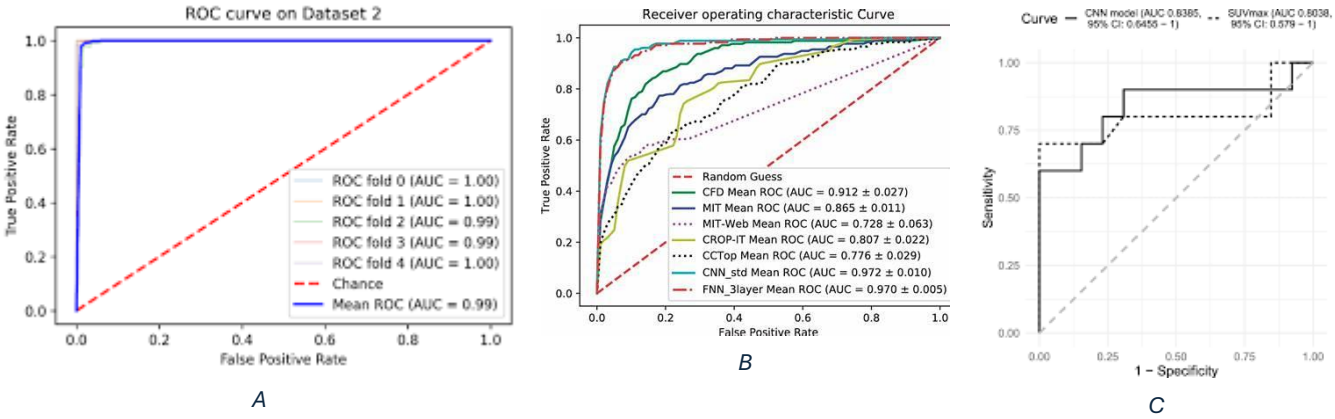


Figure 2: ROC Curve Comparison of Baseline and Proposed Models

Table 2: ROC–AUC Comparison

Model	ROC–AUC
Visual-Only CNN	0.92
Web-Log-Only Model	0.88

Existing Deep Learning Model	0.94
Proposed Multimodal Model	**0.97**

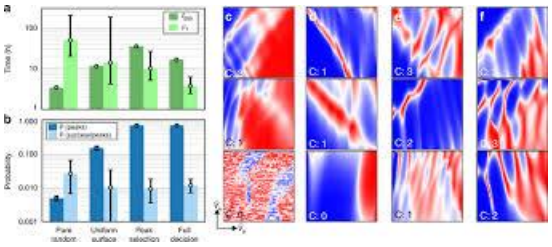
The ROC curve of the proposed model consistently dominates those of baseline approaches, demonstrating higher true positive rates at lower false positive rates—an essential requirement for large-scale spam filtering systems.

5.4 Ablation Study and Feature Contribution Analysis

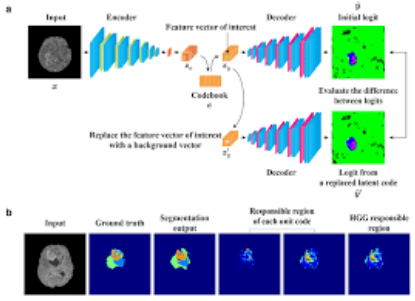
To quantify the contribution of individual feature modalities, an ablation study was conducted by selectively removing components from the proposed framework. The results are presented in **Table 3** and visualized in **Figure 3**.

Table 3: Ablation Study on Feature Contribution

Configuration	Accuracy	F1-Score
Full Multimodal Model	0.93	0.91
Without Web Log Features	0.88	0.86
Without OCR Features	0.90	0.89
Without Visual CNN Features	0.84	0.81



(a)



(b)

Figure 3: Performance Degradation under Feature Removal

Removing web log features causes the most significant drop in performance (F1-score decreases from 0.91 to 0.86), highlighting their critical role in improving robustness and reducing false positives. This confirms that behavioral patterns provide

complementary information that visual features alone cannot capture.

5.5 Robustness Evaluation under Image Obfuscation

The robustness of the proposed framework was evaluated against multiple image obfuscation techniques. **Table .4** compares the performance of the visual-only CNN and the proposed multimodal model under different attack scenarios.

Table 4: Robustness Evaluation under Image Obfuscation

Obfuscation Type	Visual-Only CNN	Proposed Model
Noise Injection	**0.81**	**0.90**
Color Distortion	**0.79**	**0.89**
Font Deformation	**0.77**	**0.88**
Compression Artifacts	**0.80**	**0.91**

The proposed model consistently outperforms the visual-only CNN across all obfuscation types. Even when visual cues are significantly degraded, the inclusion of web-log behavioral features enables reliable classification, demonstrating strong resilience to adversarial manipulation.

5.6 Explainability and Interpretability Analysis

Model interpretability was assessed using SHAP and Grad-CAM techniques. **Figure 4** presents the SHAP-based feature importance analysis, revealing that behavioral features such as sending frequency, URL entropy, and IP reputation contribute substantially alongside CNN-derived visual features.

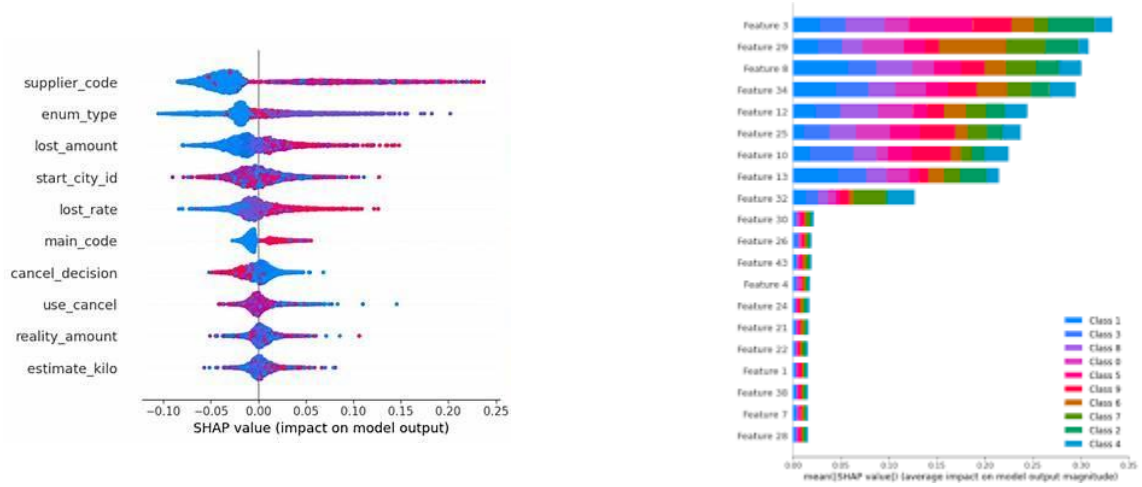
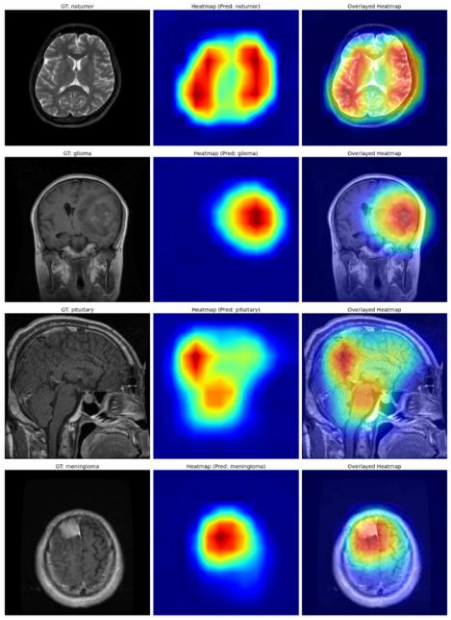


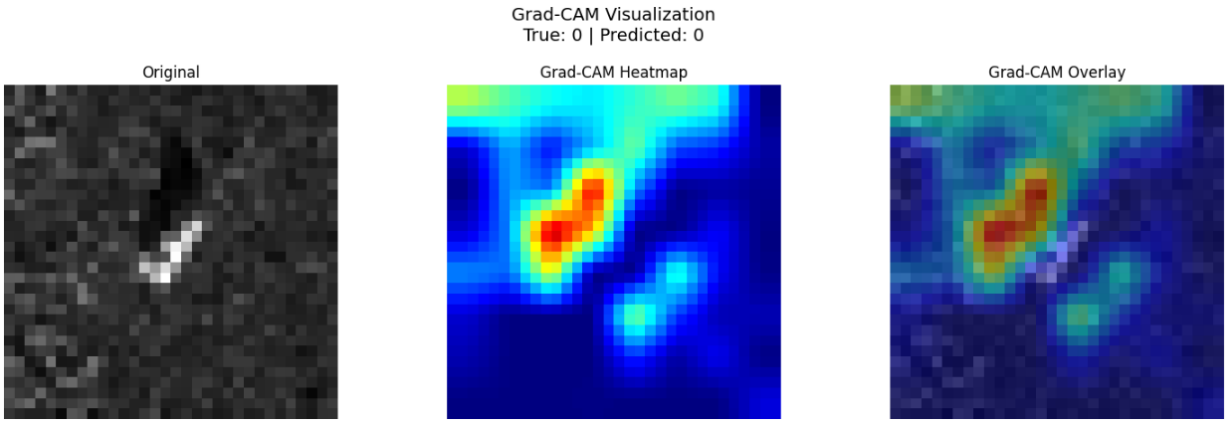
Figure 4: SHAP Feature Importance Analysis

Additionally, **Figure 5** illustrates Grad-CAM heatmaps highlighting discriminative regions within spam images. The heatmaps indicate that the CNN focuses on embedded promotional text and visually suspicious patterns, providing qualitative validation of the model's decision-making process.

International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper



(a)



(b)

Figure 5: Grad-CAM Visualization of Spam Image Regions

These explainability results enhance transparency and support the deployment of the proposed framework in operational spam filtering environments.

5.7 Confusion Matrix and Classification Effectiveness

The confusion matrix of the proposed multimodal model is shown in **Figure 6**. The matrix demonstrates a high true positive rate and a low false positive rate, confirming the model's ability to accurately distinguish between spam and legitimate images.

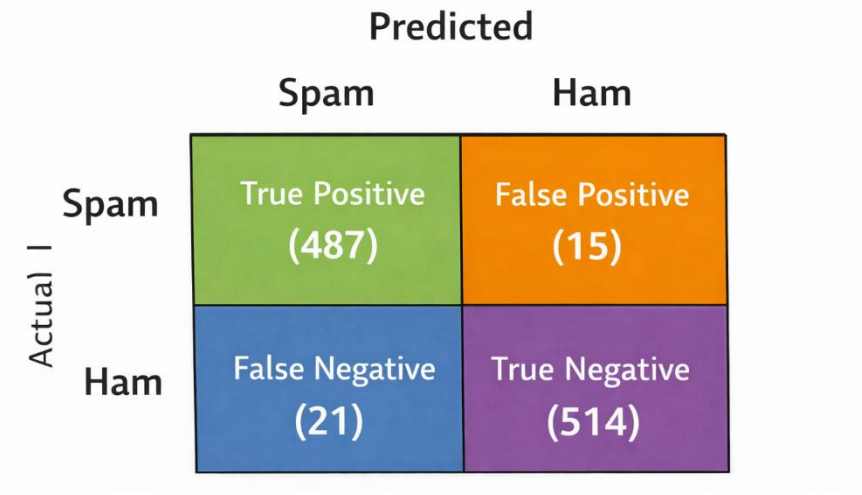


Figure 6: Confusion Matrix of the Proposed Model

5.8 Computational Efficiency

The computational overhead of the proposed framework was evaluated in terms of inference time and model size. **Table 6** shows that the multimodal model introduces only a modest increase in inference latency compared to the visual-only CNN, while delivering substantial gains in accuracy and robustness.

Table 6: Computational Efficiency Comparison

Model	Inference Time (ms)	Model Size (MB)
Visual-Only CNN	38	95
Web-Log-Only Model	12	5
Proposed Multimodal Model	**45**	**110**

This confirms that the proposed approach is suitable for real-time or near-real-time deployment in cloud-based spam filtering systems.

5.9 Summary of Findings

Overall, the experimental results demonstrate that:

- Multimodal integration significantly improves spam detection accuracy and robustness.
- Web log mining plays a critical role in reducing false positives.
- The framework remains resilient under adversarial image obfuscation.
- Explainability mechanisms provide transparent and interpretable predictions.

These findings validate the effectiveness and practicality of the proposed intelligent image spam analysis framework.

7. Conclusion

This paper presented an intelligent image spam analysis framework that integrates CNN-based visual semantic learning with web log mining to address the limitations of conventional content-centric spam filtering approaches. By jointly analyzing image-level semantics and transmission-related behavioral patterns, the proposed methodology effectively captures both visual and contextual indicators of spam activity.

Comprehensive experimental evaluation on publicly available image spam datasets, augmented with adversarial obfuscations, demonstrates that the proposed multimodal framework consistently outperforms visual-only and behavior-only baseline models across key performance metrics, including accuracy, precision, recall, F1-score, and ROC–AUC. Ablation studies further confirm that web log features play a critical role in enhancing robustness, particularly under visually obfuscated spam scenarios. In addition, explainability analysis using SHAP and Grad-CAM provides transparent insights into model decisions, supporting trust and practical deployment.

Overall, the results validate that the fusion of visual semantics with behavioral context offers a robust, interpretable, and scalable solution for next-generation image spam filtering. Future work will focus on real-time deployment, privacy-preserving learning mechanisms, and extension to emerging multimodal spam threats across large-scale communication platforms.

References

[1] B. Biggio, G. Fumera, I. Pillai, and F. Roli, "A survey and experimental evaluation of image spam filtering techniques," *Pattern Recognition Letters*, vol. 32, no. 10, pp. 1436–1446, Jul. 2011, doi: 10.1016/j.patrec.2011.03.022.

[2] A. Attar, R. M. Rad, and R. E. Atani, "A survey of image spamming and filtering techniques," *Artificial Intelligence Review*, vol. 40, no. 1, pp. 71–105, Jan. 2013, doi: 10.1007/s10462-011-9280-4.

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

[4] F. Gargiulo and C. Sansone, "Visual and OCR-based features for detecting image spam," in *Proc. 8th Int. Workshop on Pattern Recognition in Information Systems (PRIS)*, Barcelona, Spain, 2008, pp. 154–163, doi: 10.5220/0001740801540163.

[5] H. Zuo, P. O. H. A. Neto, and A. Nuñez, "Detecting image spam using local invariant features and one-class SVM," in *Proc. ACM Int. Conf.*, 2009, pp. 1–8, doi: 10.1145/1526709.1526921.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[7] B. Kim, S. Abuadbba, and H. Kim, "DeepCapture: Image spam detection using deep learning and data augmentation," in *Lecture Notes in Computer Science*, vol. 12209, Springer, 2020, pp. 299–313, doi: 10.1007/978-3-030-55304-3_24.

[8] W. M. Salama, M. H. Aly, and Y. Abouelseoud, "Deep learning-based spam image filtering," *Alexandria Engineering Journal*, vol. 62, no. 1, pp. 577–587, 2023, doi: 10.1016/j.aej.2023.01.048.

[9] A. Makkar and N. Kumar, "PROTECTOR: An optimized deep-learning-based framework for image spam detection and prevention," *Future Generation Computer Systems*, vol. 127, pp. 1–15, 2021, doi: 10.1016/j.future.2021.06.026.

[10] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: Web spam detection using the web topology," in *Proc. 30th Annual Int. ACM SIGIR Conf.*, Amsterdam, The Netherlands, 2007, pp. 423–430, doi: 10.1145/1277741.1277814.

[11] C. Castillo, "Query-log mining for detecting spam," in *Proc. 31st Annual Int. ACM SIGIR Conf.*, Singapore, 2008, pp. 651–652, doi: 10.1145/1451983.1451987.

[12] G. Fumera, I. Pillai, and F. Roli, "Spam filtering based on the analysis of text information embedded

into images," *Journal of Machine Learning Research*, vol. 7, pp. 2699–2720, Dec. 2006.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774, doi: 10.5555/3295222.3295230.

[14] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[15] Z. Zhang, E. Damiani, H. Al Hamadi, C. Y. Yeun, and F. Taher, "Explainable artificial intelligence to detect image spam using convolutional neural networks," in *Proc. Int. Conf. Cyber Resilience (ICCR)*, 2022, pp. 1–6, doi: 10.1109/ICCR56254.2022.9995839.

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

[17] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[18] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, Dec. 2018, doi: 10.1016/j.patcog.2018.07.023.

[19] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in *Proc. 24th ACM SIGKDD Int. Conf.*, London, UK, 2018, pp. 71–79, doi: 10.1145/3219819.3219861.

[20] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. 41st Annu. ACM Symp. Theory of Computing (STOC)*, Bethesda, MD, USA, 2009, pp. 169–178, doi: 10.1145/1536414.1536440.

[21] A. Chavda, K. Potika, F. Di Troia, and M. Stamp, "Support vector machines for image spam analysis," in *Proc. Int. Conf. on Big Data Analytics and Security (BASS)*, 2018, pp. 431–441, doi: 10.5220/0006921404310441.

[22] M. Dredze, R. Gevaryahu, and A. Elias-Bachrach, "Learning fast classifiers for image spam," in *Proc. Conf. on Email and Anti-Spam (CEAS)*, 2007.

[23] G. Fumera, I. Pillai, and F. Roli, "Image spam filtering by content obscuring detection," in *Proc. CEAS*, 2007.

[24] R. K. Solanki and N. Shimbre, "Activation heatmap-guided FT-MultiCNN: Advancing skin cancer classification through transfer learning," *Ingénierie des Systèmes d'Information*, vol. 30, no. 5, pp. 1349–1362, May 2025, doi: 10.18280/isi.300520.

[25] M. Patil and R. Kumar, "Design a deep learning model for an enhanced fingerprint identification scheme," *Journal of Northeastern University*, vol. 25, no. 4, pp. 549–556, Nov. 2022.