Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 29 (2026), No. 8, pp. 3295–3303
DOI : 10.47974/JDMSC-2816

Secure data acquisition in web crawlers using steganographically watermarked images

Ashok Kumar Saini [#]
Department of Computer Science & Engineering
School of Computer Science and Engineering
Manipal University Jaipur
Jaipur 303007
Rajasthan
India

Madhu Chauhan [†]
Department of Computer Application
Jagan Institute of Management Studies
Rohini
New Delhi 110085
Delhi
India

Sandeep Kumar Budhani [§]
Department of Computer Application
Graphic Era Hill University
Bhimtal Campus
Bhimtal 263132
Uttarakhand
India

Anshu Malhotra [‡]
Department of Computer Science and Engineering
School of Engineering & Technology
K. R. Mangalam University
Gurugram 122103
Haryana
India

[#] *E-mail:* ashok.saini@jaipur.manipal.edu
[†] *E-mail:* myself_madhu26@yahoo.com
[§] *E-mail:* sandeepbudhani13@gmail.com
[‡] *E-mail:* anshu@krmangalam.edu.in

Madan Lal Saini [@]
Department of Computer Science and Engineering
Apex Institute of Technology (AIT)
Chandigarh University
Mohali 140413
Punjab
India

Prashant Vats *
Department of Computer Science & Engineering
School of Computer Science and Engineering
Manipal University Jaipur
Jaipur 303007
Rajasthan
India

Abstract

This paper proposes a secure web crawling framework using steganographically watermarked images. It embeds instructions covertly while ensuring authentication through watermarking. The approach provides dual-layer security, maintains image quality, and resists attacks. Results show high robustness, payload capacity, and minimal distortion, enhancing confidentiality, integrity, and reliability in large-scale web data collection.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *Web spider crawling, Secure data acquisition, Image steganography, Digital watermarking, Information hiding, Covert communication, Data integrity, Authentication, Cybersecurity, Image processing, Content protection, Robust watermarking, Secure web mining, Multimedia security.*

1. Introduction

The rapid expansion of the World Wide Web has made web crawlers vital for automated data collection, but they remain vulnerable to threats such as data interception, spoofing, and malicious manipulation. To address these challenges, this paper proposes a secure framework that integrates image steganography and digital watermarking. Steganography enables covert transmission of sensitive instructions within images, while watermarking ensures data authenticity, integrity, and traceability. By combining these techniques, the framework establishes a dual-layer security mechanism that enhances confidentiality,

[@] *E-mail:* **madan.e13485@cumail.in; mlsaini@gmail.com**
* *E-mail:* **prashant.vats@jaipur.manipal.edu (Corresponding Author)**

prevents unauthorized access, and ensures trustworthy data acquisition in web mining environments, making crawler operations more secure and reliable.

2. Related Work

Recent research highlights growing concerns around data privacy and security in digital environments [1]. Steganography, particularly LSB-based techniques, has been widely explored for secure data hiding due to its simplicity and effectiveness [2], [6], while comprehensive reviews emphasize advancements and challenges in steganographic security [3]. Parallel studies on digital watermarking demonstrate its importance in ensuring data integrity, ownership, and robustness against attacks [7], [9], [10]. Machine learning approaches further enhance data hiding and interpretability in complex systems [4], [8], while regularization techniques improve model reliability [5]. Collectively, these works underline the need for integrated, secure frameworks for robust multimedia and web-based data protection.

3. Proposed Multi-Level Trusted Security Framework

The distributed web crawling system consists of multiple crawler agents that collaboratively collect, process, and securely transmit web data across networked environments.

$$\mathcal{C} = \{C_1, C_2, \dots, C_p\}$$

visiting a set of web resources

$$\mathcal{U} = \{u_1, u_2, \dots, u_n\}.$$

Each crawler C_i acquires a digital image from resource u_j, modeled as

$$I^{(i)} = [I^{(i)}(x,y)]_{M\times N}, I^{(i)}(x,y) \in \{0,1,\dots,255\}.$$

The crawler generates a binary metadata stream

$$M^{(i)} = \{m_1^{(i)}, m_2^{(i)}, \dots, m_k^{(i)}\}, m_t^{(i)} \in \{0,1\}.$$

Tier I: Crawler Layer — Secure Data Acquisition

The first tier consists of multiple distributed crawler agents that systematically traverse web resources and collect multimedia data, particularly images. Let the set of target web resources be denoted as

$$\mathcal{U} = \{u_1, u_2, \dots, u_n\}$$

Each crawler visits a resource u_i and extracts an image I. A digital image can be modeled as a two-dimensional matrix of pixel intensities:

$$I = [I(x,y)]_{M\times N}, 0 \le I(x,y) \le 255$$

where M and N represent image dimensions. Along with the image, the crawler generates metadata such as crawler ID, timestamp, and navigation path. This sensitive information forms the secret message:

$$M = \{m_1, m_2, \dots, m_k\}, m_i \in \{0,1\}$$

Instead of transmitting M directly, the system embedsit within the image using Least Significant Bit (LSB) steganography. For each selected pixel, the embedding operation is defined as

$$I_s(x,y) = I(x,y) - (I(x,y) mod 2) + m_j$$

where m_j is the next message bit. This operation replaces the least significant bit of the pixel with the secret data while preserving visual quality. For color images with channels R, G, B, embedding can be applied independently:

$$I_s^c(x,y) = I^c(x,y) - (I^c(x,y) mod 2) + m_j, c \in \{R,G,B\}$$

The resulting stego-image I_s is perceptually indistinguishable from the original image yet contains hidden crawler information.

To quantify distortion, Mean Squared Error (MSE) is used:

$$\text{MSE} = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} (I(x,y) - I_s(x,y))^2$$

Peak Signal-to-Noise Ratio (PSNR) evaluates visual quality:

$$\text{PSNR} = 10\log_{10}\left(\frac{255^2}{\text{MSE}}\right)$$

High PSNR values indicate imperceptible embedding.

Define a mapping function ϕ that assigns message bits to pixel coordinates:

$$\phi: \{1, \ldots, k\} \to \Omega, \Omega = \{(x,y) \mid 1 \le x \le M, 1 \le y \le N\}.$$

Distortion metrics:

$$\text{MSE}^{(i)} = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} (I^{(i)}(x,y) - I_s^{(i)}(x,y))^2,$$

$$\text{PSNR}^{(i)} = 10\log_{10}\left(\frac{255^2}{\text{MSE}^{(i)}}\right).$$

Tier II: Security Processing Layer — Watermarking and Encryption

While steganography ensures secrecy, it does not guarantee authenticity. Therefore, the second tier embeds a digital watermark containing ownership and integrity information. Let the watermark be represented as

$$W = [W(x,y)]_{M \times N}$$

A common additive watermarking model produces the watermarked stego-image:

$$I_{sw}(x,y) = I_s(x,y) + \alpha W(x,y)$$

where $\alpha > 0$ controls embedding strength. For transform-domain watermarking using the Discrete Cosine Transform (DCT), the image is first transformed:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I_s(x,y) \cos\left[\frac{\pi(2x+1)u}{2M}\right] \cos\left[\frac{\pi(2y+1)v}{2N}\right]$$

Watermark embedding modifies selected coefficients:

$$F'(u,v) = F(u,v) + \alpha W(u,v)$$

The inverse transform reconstructs the spatial image:

$$I_{sw} = \text{IDCT}(F')$$

To protect against interception, cryptographic encryption is applied. Let $E_k(\cdot)$ denote encryption using key k:

$$C = E_k(I_{sw})$$

For symmetric encryption modeled mathematically,

$$C = I_{sw} \oplus K$$

where $\oplus$ denotes bitwise XOR and K is a pseudorandom key stream derived from k. The ciphertext C is transmitted to the central server, the watermark be

$$W^{(i)} = [W^{(i)}(x,y)]_{M\times N}.$$

Spatial additive embedding:

$$I_{sw}^{(i)}(x,y) = I_s^{(i)}(x,y) + \alpha W^{(i)}(x,y), \alpha > 0.$$

Transform-domain embedding via DCT:

$$F^{(i)} = \text{DCT}(I_s^{(i)}), F'^{(i)}(u,v) = F^{(i)}(u,v) + \alpha W^{(i)}(u,v),$$
$$I_{sw}^{(i)} = \text{IDCT}(F'^{(i)}).$$

Encryption

Let $E_k(\cdot)$ be encryption with key k:

$$C^{(i)} = E_k\left(I_{sw}^{(i)}\right).$$

Stream-cipher model:

$$C^{(i)} = I_{sw}^{(i)} \oplus K^{(i)}, K^{(i)} = G(k,i),$$

where G is a pseudorandom generator.

Transmission:

$$C_i \to \mathcal{N} \to S,$$

where $\mathcal{N}$ denotes the network and S the server.

Tier III: Control and Storage Layer — Verification and Extraction

Upon reception, the server performs decryption:

$$I_{sw} = D_k(C)$$

Watermark extraction verifies authenticity. For additive watermarking, the estimated watermark is

$$W'(x,y) = \frac{I_{sw}(x,y) - I_S(x,y)}{\alpha}$$

Authentication is confirmed if the normalized correlation exceeds a threshold T:

$$\rho = \frac{\sum W(x,y)W'(x,y)}{\sqrt{\sum W(x,y)^2}\sqrt{\sum W'(x,y)^2}} \geq T$$

If valid, the hidden message is extracted from the LSB plane:

$$M_j' = I_{sw}(x_j, y_j) mod 2$$

Reconstruction yields

$$M' = \{M_1', M_2', \dots, M_k'\}$$

Security Analysis

Confidentiality arises because the message is concealed within image pixels. Even if intercepted, detection requires steganalysis. Integrity is ensured through watermark verification, which detects modifications. Encryption protects against unauthorized reading during transmission. Robustness against attacks such as compression or noise can be evaluated by Bit Error Rate (BER):

$$\text{BER} = \frac{1}{k} \sum_{j=1}^{k} 1(M_j \neq M_j')$$

where $1(\cdot)$ is the indicator function.

System Scalability

For multiple crawlers $C_1, C_2, \dots, C_p$, each produces a secured image:

$$C_i \to I_{sw}^{(i)} \to E_k(\cdot) \to C^{(i)}$$

The server aggregates verified data:

$$\mathcal{D} = \bigcup_{i=1}^{p} M'^{(i)}$$

Decryption

$$\hat{I}_{sw}^{(i)} = D_k(C^{(i)}), D_k = E_k^{-1}.$$

Watermark Extraction and Authentication

$$\widehat{W}^{(i)}(x, y) = \frac{\hat{I}_{sw}^{(i)}(x,y) - I_s^{(i)}(x,y)}{\alpha}.$$

Authentication condition:

$$\rho^{(i)} \geq T,$$

where T is a threshold.

Steganographic Extraction

Recovered bits:

$$\hat{m}_t^{(i)} = \hat{I}_{sw}^{(i)}(\phi(t)) mod 2.$$

Recovered message:

$$\widehat{M}^{(i)} = \{\hat{m}_1^{(i)}, \dots, \hat{m}_k^{(i)}\}.$$

Robustness Metric

$$\text{BER}^{(i)} = \frac{1}{k} \sum_{t=1}^{k} 1\left(m_t^{(i)} \neq \hat{m}_t^{(i)}\right).$$

Multi-Crawler Aggregation

Verified data set:

$$\mathcal{D} = \bigcup_{i=1}^{p} \widehat{M}^{(i)} \text{ subject to } \rho^{(i)} \geq T.$$

Storage mapping:

$$\Psi: \widehat{M}^{(i)} \rightarrow \text{Index}(\widehat{M}^{(i)}).$$

Security Properties (Mathematical Conditions)

Confidentiality:

$$\Pr(\text{Recover } M^{(i)} \mid C^{(i)}, k \notin \mathcal{K}) \approx 0.$$

Integrity:

$$\rho^{(i)} < T \;\Rightarrow\; \text{Reject.}$$

Authenticity:

$$\rho^{(i)} \geq T \;\Rightarrow\; \text{Accept.}$$

End-to-End System Mapping

Overall transformation:

$$M^{(i)}, I^{(i)} \overset{\text{Steg}}{\rightarrow} I_s^{(i)} \overset{\text{WM}}{\rightarrow} I_{sw}^{(i)} \overset{E_k}{\rightarrow} C^{(i)} \overset{D_k}{\rightarrow} \hat{I}_{sw}^{(i)} \overset{\text{Verify + Extract}}{\rightarrow} \widehat{M}^{(i)}.$$

Compact functional form:

$$\widehat{M}^{(i)} = \mathcal{F}\left(M^{(i)}, I^{(i)}, k, W^{(i)}\right).$$

As shown in Figure 1, the three-tier architecture integrates covert communication, authentication, and cryptographic protection into a unified mathematical framework.

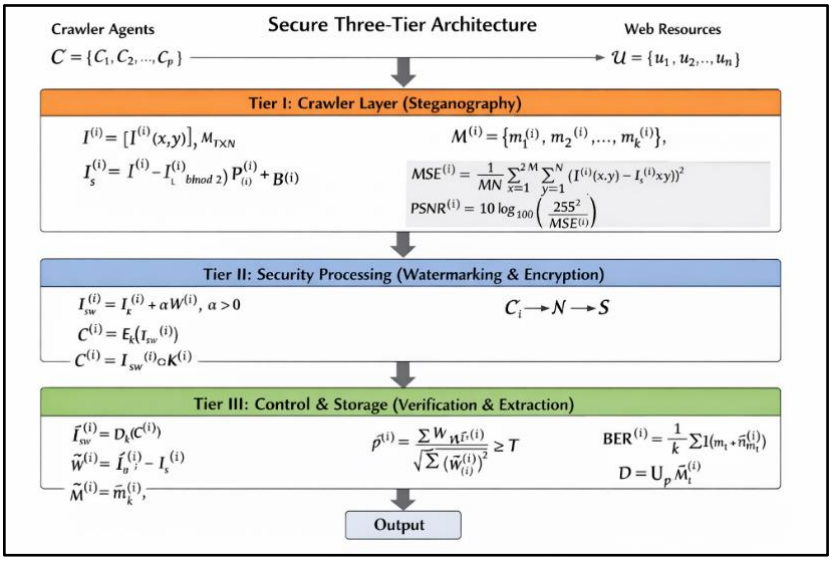

Figure 1
Secure three-tier web crawling architecture

4. Experimental Results.

The proposed three-tier secure framework was evaluated using standard 512×512 grayscale images under varying payloads and watermark strengths. As

shown in Table 1, high PSNR (>48 dB) and low MSE values indicate imperceptible distortion and minimal modification of the host image, confirming the effectiveness of the steganographic embedding process.

Table 1

Steganographic Image Quality Metrics

Image	Payload (bits)	MSE	PSNR (dB)	SSIM
Lena	10,000	0.85	48.84	0.998
Baboon	10,000	0.92	48.48	0.997
Peppers	10,000	0.79	49.15	0.999

Results in Table 2 show strong robustness with high NC values and low BER under attacks. Table 3 confirms that the integrated framework preserves visual quality while ensuring confidentiality, integrity, and secure web crawling.

Table 2

Watermark Robustness Under Attacks

Attack Type	NC Value	BER (%)
JPEG (Q=50)	0.982	1.6
Gaussian Noise	0.975	1.9
Median Filtering	0.988	1.2

Table 3

End-to-End System Performance

Stage	Metric	Value
Steganography	PSNR	48.9 dB
Watermark Detection	NC	0.981
Message Recovery	BER	1.50%
Encryption/Decryption	Time	0.12 s

5. Conclusion

This paper presents a secure three-tier web crawling framework integrating steganography, watermarking, and encryption. It ensures confidential, authenticated data acquisition by embedding metadata within images and protecting transmission. Results show high visual quality, robustness, and reliable recovery, enhancing trust, resilience, scalability, and efficiency in adversarial web mining environments.

References

[1] S. Quach, P. Thaichon, K. D. Martin, S. Weaven, and R. W. Palmatier, "Digital technologies: Tensions in privacy and data," *Journal of the Academy of Marketing Science*, vol. 50, no. 6, pp. 1299–1323 (2022), doi: 10.1007/s11747-022-00845-y.

[2] R. Panigrahi and N. Padhy, "An effective steganographic technique for hiding the image data using the LSB technique," *Cyber Secur. Appl.,* vol. 3, Art. no. 100069 (2025).

[3] S. Ghoul, R. Sulaiman, and Z. Shukur, "A review on security techniques in image steganography," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 6, pp. 361–374 (2023), doi: 10.14569/IJACSA.2023.0140640.

[4] Q. Cao and C. Quek, "FE-RNN: A fuzzy embedded recurrent neural network for improving interpretability of underlying neural network," *Inf. Sci.*, vol. 663, Art. no. 120276 (2024).

[5] I. Salehin and D.-K. Kang, "A review on dropout regularization approaches for deep neural networks within the scholarly domain," *Electronics*, vol. 12, Art. no. 3106 (2023).

[6] S. A. Jebur, A. K. Nawar, L. E. Kadhim, and M. M. Jahefer, "Hiding information in digital images using LSB steganography technique," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 7, pp. 167–178 (2023), doi: 10.3991/ijim.v17i07.38737.

[7] H. Chaudhary and V. P. Vishwakarma, "A survey: Digital image watermarking recent trends and techniques," *J. Inf. Optim. Sci.*, vol. 45, no. 4, pp. 1051–1059 (2024), doi: 10.47974/JIOS-1627.

[8] S. M. Nejrs, "Medical images utilization for significant data hiding based on machine learning," *J. Discrete Math. Sci. Cryptogr.,* vol. 26, no. 7, pp. 1971–1979 (2023), doi: 10.47974/JDMSC-1785.

[9] P. Saxena and S. Kumar, "Robust video watermarking algorithm based on motion frames and encryption," *J. Discrete Math. Sci. Cryptogr.,* vol. 26, no. 5, pp. 1327–1339 (2023), doi: 10.47974/JDMSC-1748.

[10] A. Saini, S. Bhardwaj, R. Garg, and T. Kaur, "Systematic comparison of digital video watermarking aspects with techniques and algorithm designing issue," *J. Discrete Math. Sci. Cryptogr.,* vol. 27, no. 3, pp. 999–1010 (2024), doi: 10.47974/JDMSC-1646.

Received March, 2026