Journal of Soft Computing Exploration
Homepage: htttps://shmpublisher.com/index.php/joscex
p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

An integrated email security gateway and dual-layer frequency-domain watermarking framework for business transaction receipt authentication against generative AI manipulation

Bonifacius Vicky Indriyono[1], Rabei Raad Ali[2]
[1]Department of Digital Business, Universitas STRADA Indonesia, Kediri, Indonesia
[2]Artificial Intelligence Research Center, Northern Technical University, Iraq

Article Info

Article history:

Received April 30, 2026
Revised June 16, 2026
Accepted July 4, 2026

Keywords:

Digital watermarking
Document fraud
Generative artificial intelligent
Email security
Tamper detection

ABSTRACT

Business document forgery and email phishing have been significantly amplified by generative artificial intelligence, yet existing security frameworks address communication-channel protection and document authentication as separate concerns, leaving organizations vulnerable to coordinated multi-vector attacks. This research aims to design, implement, and evaluate a unified "Dual-Shield" framework that bridges this gap through an integrated, lightweight solution for real-time business operations. The proposed system combines a four-stage email security gateway—covering sender verification, URL analysis via VirusTotal API, and attachment sanitization—with a dual-layer digital watermarking engine applied to transaction receipt images. The watermarking component pairs a robust Discrete Wavelet Transform–Singular Value Decomposition layer for authorship attribution with a semi-fragile Quantization Index Modulation–Discrete Cosine Transform layer for monetary integrity verification, assessed via imperceptibility, robustness, and tamper-detection metrics on synthetic and real-world datasets. Results show the robust watermark achieved peak signal-to-noise ratio values of 43.38–43.48 dB with SSIM above 0.98. All generative AI manipulations were detected with bit error rates of 0.375–0.472, and targeted ten-percent numerical fraud was identified by the semi-fragile layer. The gateway correctly classified all test scenarios, confirming that malicious URL detection overrides trusted sender status to prevent whitelist exploitation.

Corresponding Author:

Bonifacius Vicky Indriyono,
Department of Digital Business
Universitas STRADA Indonesia
Manila Street No.37, Tosaren, Kediri, East Java 64133
Email: bonifaciusvicky@gmail.com
https://doi.org/10.52465/joscex.v7i3.77

1. INTRODUCTION

The rapid growth of digital transformation has enhanced business efficiency while increasing vulnerability to financial fraud. Phishing remains among the most prevalent cybercrimes, with incidents nearly doubling since 2020 [1]. The seriousness of this threat is reflected in the FBI's Internet Crime Complaint Center (IC3) report, which recorded more than 22,000 AI-enabled fraud complaints in 2025, resulting in losses

exceeding $893 million, including over $30 million from AI-assisted Business Email Compromise (BEC) scams [2]. Generative AI has further expanded the attack surface by automating sophisticated attacks such as credential theft, document forgery, phishing, and malware distribution [3]. Modern phishing campaigns employ highly personalized AI-generated messages impersonating trusted entities, while malicious links can deploy ransomware, keyloggers, and credential-stealing scripts [1], [3]. VirusTotal-based machine-learning approaches have proven more effective than majority-voting methods for detecting phishing and malware-hosting URLs [4], motivating its adoption as the URL analysis component in this study. Since phishing often precedes document fraud, this research integrates email threat detection with document authentication to secure both communication channels and transaction document integrity.

In Indonesia, the need for such integration is underscored by cyber incidents reported by BSSN, which rose from 290,000 cases in 2019 to 1,031,389 in 2023 [5]. In 2025, BSSN recorded 3.64 billion traffic anomalies within seven months, while ransomware attacks such as LockBit 3.0 and Brain Cipher disrupted more than 200 public services [6]. Simultaneously, manipulation of digital transaction evidence, including modified invoices and fraudulent receipts, threatens corporate financial integrity. These multi-vector threats emphasize the necessity of a dual-protection system capable of ensuring both communication security and document authenticity within a unified workflow.

Although digital watermarking has advanced document integrity protection, existing solutions remain fragmented and limited in practice. Robust watermarking techniques, such as the randomized smoothing framework by Jiang et al., provide resistance against watermark removal and AI-assisted manipulation, achieving near-zero false negatives under JPEG attacks with an average SSIM of 0.943 [7]. However, these approaches focus solely on document integrity and neglect communication security. Other techniques also present drawbacks: the DWT-DCT method achieves high visual quality (PSNR up to 57 dB) but performs poorly under severe JPEG compression [8], the DWT-SVD-QR-RSA scheme is robust yet computationally expensive and visually intrusive [9], Naffouti et al.'s non-blind DWT-SVD method requires original data during extraction [10]. CNN-based approaches demand extensive computational resources and large datasets [11]. Likewise, phishing detection frameworks can achieve high accuracy (99.85%) but remain resource-intensive and disconnected from document authentication processes [12]. Most studies prioritize watermark imperceptibility while overlooking compression robustness and integration with email security. To address these gaps, this study proposes a lightweight Dual-Shield framework that combines a four-stage cryptographic pipeline with dual watermarking, integrating robust watermarking for authorship attribution and semi-fragile watermarking for transactional integrity. The framework also incorporates an automated email security gateway to inspect communications before document processing, bridging the longstanding separation between communication security and document validation amid increasingly sophisticated BEC attacks that circumvent traditional mechanisms such as DMARC, SPF, and DKIM after breaching organizational perimeters [13].

Existing forgery detection methods remain insufficient against sophisticated deep learning-based document manipulations, such as advanced inpainting and scene text editing that leave no visible traces [11]. To address this, this study proposes a two-stage defense mechanism integrating network threat intelligence and transform-domain cryptography while preserving computational efficiency through spatial-domain watermark embedding. Combined with a non-blocking email gateway, the proposed watermarking strategy enables secure and automated business communication, resilience to lossy transmission, effective tampering detection, and low processing latency [12]. This integrated framework provides a comprehensive, computationally efficient anti-fraud solution for real-time business administration by simultaneously ensuring document authenticity and communication security, as summarized in Table 1.

Table 1. Research gap analysis matrix of the proposed system against prior works

Reference	Methodology	Unresolved Issue
Papathanasiou et al. (2024) [13]	QR Code methodology & email authentication protocols	Completely blind to post-delivery text manipulations or generative visual modifications inside administrative attachments.
Okamoto et al. (2023) [11]	Self-supervised learning & DNN forensics classifiers	Passive detection only; highly effective for post-fraud forensics but incapable of active perimeter prevention or live channel blocking.
Zhang & Su (2021) [12]	Spatial domain embedding via multilevel DHT	Focuses strictly on standalone image authentication; ignores upstream network vulnerabilities and transmission channel security.

2. METHOD

The proposed system adopts a dual-watermarking strategy that embeds robust watermarks for authorship verification and fragile watermarks for transactional integrity validation [6], [7] as illustrated I nFigure 1 and Table 2. As the first defense layer, the email security gateway module (email_gateway.py) executes a four-stage threat detection pipeline [8] : (1) Email Parsing extracts sender information, hyperlinks, and attachments from .eml files using RFC 5322-compliant text-based processing to maintain real-time efficiency [14]; (2) Sender Verification checks senders against an SQLite customer database to identify unregistered or suspicious sources; (3) URL Analysis examines text-based links for typosquatting and phishing indicators [15] [16], while image-based phishing such as QR-code phishing (quishing) falls outside the gateway scope and is mitigated by the Dual-Watermarking layer protecting financial document attachments [14]; and (4) Attachment Sanitization validates extensions and MIME types to ensure only legitimate image formats are accepted for e-receipt generation [17]. URLs are classified using VirusTotal multi-vendor aggregations into Malicious, Suspicious, and Clean categories. A URL is considered Malicious if detected by at least one security vendor (≥1 detection), following a zero-trust approach that prioritizes early interception and minimizes false negatives [18]. Suspicious covers ambiguous cases involving untrusted vendors or inconsistent early detections, addressing delays in identifying emerging phishing domains [18]. A URL is labeled Clean only when it receives zero detections across all VirusTotal scanners, providing the highest confidence before permitting traffic through the security gateway [18].

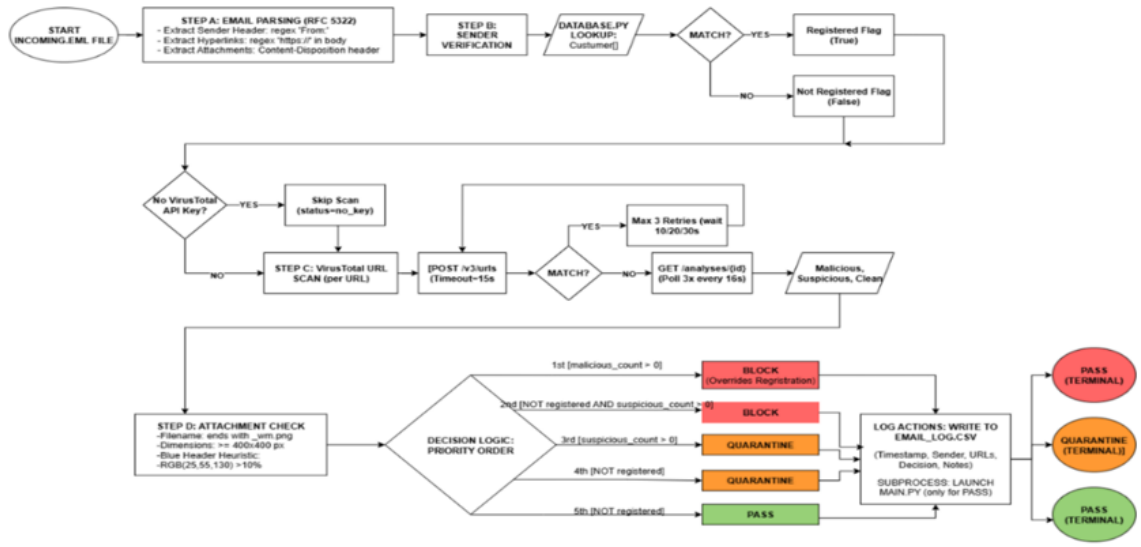

Figure 1. Research workflow

Table 2. Email gateway decision matrix

Decision	Trigger Condition	System Action
BLOCK	≥1 malicious URL, or unknown sender + suspicious URL	Email discarded; application not launched
QUARANTINE	≥1 suspicious URL, or unknown sender without malicious URL	Email held for manual review
PASS	Registered sender, no malicious or suspicious URLs	main.py launched; verification proceeds

Receipt Image Generation

The receipt generator (receipt_generator.py) constructs a 600-pixel-wide PNG image on a white (RGB 255,255,255) background. White maximises perceptual invisibility for DWT-SVD watermarking [19]. Image dimensions are constrained to multiples of four pixels using equation (1), where This satisfies the alignment requirement of a two-level DWT decomposition, which requires dimensions divisible by 2^level = 4 [20]. Violating this constraint introduces boundary artefacts in the sub-band coefficients.

$$img.crop(\ (0, 0, (W//4) \times 4, (H//4) \times 4)\ ) \quad (1)$$

Dual Watermarkign Engine

Colour space conversion

All watermarking operations were performed on the luminance (Y) channel obtained through YCbCr conversion in Equations (2)–(4). The Y channel was selected because JPEG compression applies chroma subsampling to the Cb and Cr channels, making watermarks embedded in chrominance components highly susceptible to degradation under recompression. Although the human visual system is more sensitive to luminance variations than chrominance changes [20] , [21], this perceptual trade-off was mitigated by

embedding the watermark only within the LL2 sub-band of a two-level DWT decomposition, where low-frequency modifications are less perceptible [20], and by carefully limiting the embedding strength (α) to preserve visual quality. Experimental results validate this balance, achieving PSNR values of 43.38–43.48 dB and SSIM values above 0.98, both exceeding the imperceptibility thresholds of 40 dB and 0.95, respectively.

$$Y = 0.299\,R + 0.587\,G + 0.114\,B \tag{2}$$

$$Cb = -0.168736R - 0.331264G + 0.5B + 128 \tag{3}$$

$$Cr = 0.5R - 0.418688G - 0.081312B + 128 \tag{4}$$

Subsequently, features were extracted to produce additional attributes that are relevant for the analysis. Gestational age in days was calculated by subtracting the date of last menstrual period (HPHT) from the date of birth, and then both original date attributes were removed. Also, inter-pregnancy interval values were converted from years to months by dividing the recorded values by 12, which ensured consistency in measurement units across the dataset.

Two-level discrete wavelet transform (DWT)

The Haar wavelet basis is used for both decomposition and reconstruction. The analysis filters are illustrated in equation (5) and equation (6), where two-level decomposition produces the LL2 sub-band, which concentrates maximum image energy and is least affected by JPEG quantisation or noise addition [19][21]. This sub-band is selected as the embedding domain for the robust watermark.

$$h_L = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right] (low - pass) \tag{5}$$

$$h_H = \left[\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right] (high - \ pass) \tag{6}$$

Singular value decomposition (SVD)

SVD is applied to the LL2 sub-band matrix A as in equation (7), where U and V are orthogonal matrices and Sigma is diagonal with singular values sigma_1 >= sigma_2 >= ... >= 0. The robust watermark modifies each singular value according to the watermark bit w_i as in equation (8) and equation (9), where alpha = 10.0 (ALPHA_ROBUST). Extraction applies the sign-of-difference rule [22]. Because the embedded delta (±alpha) is much larger than perturbation from JPEG compression at quality >= 70, Bit Error Rate (BER) remains below 0.25 [19][21]. To prevent direct reading of watermark bits, the bit sequence is scrambled using a Chebyshev chaotic map as in (10). The initial seed x_0 is derived from the Message Digest (MD5) of the receipt identifier, ensuring uniqueness and exact reproducibility at verification without transmission.

$$A = U\,\Sigma\,VT \tag{7}$$

$$\sigma_i' = \sigma_i + \alpha \times (+1\ if w_i = 1, -1\ if w_i = 0) \tag{8}$$

$$w_i' = 1\ if(\sigma_i' - \sigma_i) > 0, else\ w_i' = 0 \tag{9}$$

$$x_{k+1} = cos(n \cdot arccos(x_k)), \quad x_k \in [-1,1], \quad n = 4 \tag{10}$$

QIM-DCT semi-fragile watermark

The transaction total is embedded in mid-frequency DCT coefficients of 8×8 pixel blocks using Quantization Index Modulation (QIM) [23]. Given step size Delta = 30 as follow in equation (11) until (14), where extraction requires only the received coefficient c" and the known step size as illustrated in equation (15). The 72-bit payload consists of an 8-bit synchronisation marker (0xAF) followed by the transaction total encoded as a 64-bit IEEE 754 double-precision float, guaranteeing lossless monetary encoding.

$$q = round(c/\Delta) \tag{11}$$

$$q* = q + 1 \; if \; b = 1 \; and \; (q \; mod \; 2) \neq 1 \tag{12}$$

$$q* = q - 1 \; if \; b = 0 \; and \; (q \; mod \; 2) \neq 0 \tag{13}$$

$$c' = q^* \times \Delta \tag{14}$$

$$\hat{b} = round(c'' / \Delta) \; mod \; 2 \tag{15}$$

Proposed verification network

The full_verification() process performs four sequential stages to distinguish between benign recompression and malicious tampering [24][25]. As illustrated in Figure 2, Layer 1 verifies SHA-256 hash integrity by comparing the received file's hash with the original issuance hash. Since any file modification changes the digest, a failed hash check does not immediately invalidate the receipt and necessitates further analysis. Layer 2 evaluates the robust DWT-SVD watermark by extracting the receipt ID using the stored chaotic key and singular values, employing Bit Error Rate (BER) and Normalized Correlation (NC) metrics (Equations (16) and (17)). Provenance is confirmed when BER < 0.25 and NC > 0.75, allowing recompressed receipts to remain valid [20], [26]. In Layer 3, the semi-fragile QIM-DCT watermark is assessed through extraction of the 72-bit message and marker synchronization. Receipts pass when BER < 0.15, where BER < 0.05 typically indicates recompression and BER > 0.15 suggests manipulation [27]. Finally, Layer 4 validates transactional integrity by comparing the extracted 64-bit floating-point transaction amount (bits 8–71) against the stored value with a tolerance of ±0.01.

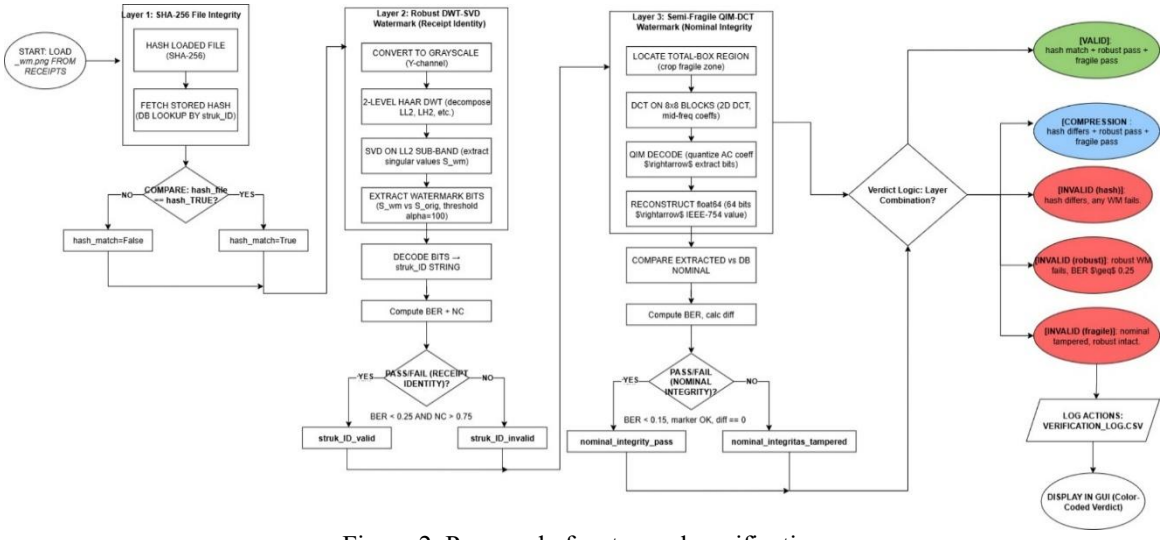

Figure 2. Proposed of watermark verification

Table 3. Combined four-stage verification outcome interpretation

Stage A	Stage B	Stage C	Stage D	Interpretation
PASS	PASS	PASS	PASS	Fully authentic; file is byte-identical to original
FAIL	PASS	PASS	PASS	File recompressed (e.g. WhatsApp/JPEG); content intact
FAIL	PASS	FAIL	FAIL	Content manipulated; financial amounts altered
FAIL	FAIL	FAIL	FAIL	Severe failure; receipt identifier unrecoverable; rejected

$$BER = \frac{\sum_{i=1}^{N} |\widehat{w_i} - w_i|}{N} \tag{16}$$

$$NC = \frac{\widetilde{w} \cdot \widehat{\widetilde{w}}}{\|\widetilde{w}\| \times \|\widehat{\widetilde{w}}\|} \tag{17}$$

Dataset

To evaluate the proposed email security gateway, this study employed both a controlled synthetic dataset and public real-world datasets. The synthetic dataset, consisting of 10 custom email scenarios, was designed to align with the internal SQLite customer database, enabling precise assessment of the gateway's priority-based decision logic, particularly Sender Verification. To examine scalability and real-world

performance, two public datasets were further utilized within the parsing and URL analysis pipeline to evaluate the VirusTotal-based detection mechanism and the gateway's classification thresholds under realistic conditions. The "Phishing Email Dataset" by Naser Abdullah Alam was used exclusively to assess Email Domain Detection in Table 6 by parsing sender addresses and domain information to challenge Step B (Sender Verification) in distinguishing legitimate communications from spoofed or malicious domains. Meanwhile, the "Malicious URLs Dataset" by Sid321axn, containing benign, phishing, malware-hosting, and defacement URLs, was employed exclusively to evaluate URL Phishing Detection in Table 6. These URLs were processed through Step C (URL Analysis) to benchmark the VirusTotal API aggregation logic and assess the effectiveness of the gateway's zero-trust threshold in identifying malicious links regardless of sender legitimacy.

3. RESULTS AND DISCUSSIONS

Watermark Quality Test Results

Five transaction receipts were generated for automated quality testing. Table 3 shown summarizes the metrics for groups T1 to T6, where overall FAIL due to T5 (salt & pepper) and T6 (rotation). Based on Table 3. the evaluation demonstrated that all receipts achieved excellent imperceptibility, with PSNR values ranging from 43.38–43.48 dB and SSIM values above 0.98, exceeding the benchmarks of 40 dB and 0.95, respectively. These findings are consistent with Utomo et al. [30], confirming that the use of a scaling factor of 100 in the DWT-SVD scheme preserves high visual quality. In terms of JPEG robustness, the system achieved BER = 0.000 at quality factors Q90 and Q80, while only a minimal BER of 0.0069 was observed at Q70, remaining well below the 0.25 safety threshold due to embedding within the resilient LL2 sub-band. For semi-fragile tamper detection (T3), the QIM-DCT layer successfully detected modifications involving price changes, total amount alterations, and combined attacks, producing BER values between 0.069 and 0.208, with most cases exceeding the 0.10 detection threshold. However, one receipt (STR-20260429011812) recorded a BER of 0.069, indicating that highly localized modifications may occasionally evade detection. The system's limitations were also identified: T5 failed because noise robustness and imperceptibility share the same parameter (α), creating an inherent trade-off in which improving one degrades the other; the chosen configuration prioritizes JPEG robustness and visual quality to align with the operational threat model. Meanwhile, T6 failed because DWT-SVD is inherently grid-aligned and lacks geometric invariance, and addressing this limitation would require synchronization mechanisms that conflict with the framework's lightweight design objective. Moreover, geometric transformations fall outside the intended threat model of a digital-file processing pipeline.

Table 3. Watermark quality test

Receipt Number	Customer	T1 PSNR (dB)	T1 SSIM	T2 Q70 BER	T3 Attack BER	T4 BPP	All Pass?
STR-20260429011542	Budi Santoso	43.48	0.9855	0.0069	0.1944	0.00045	FAIL*
STR-20260429011606	Siti Rahayu	43.45	0.9861	0.0069	0.1389	0.00045	FAIL*
STR-20260429011812	Reza Firmansyah	43.46	0.9858	0.0069	0.0694	0.00045	FAIL*
STR-20260429011910	Nurul Hidayah	43.45	0.9860	0.0000	0.1667	0.00045	FAIL*
STR-20260429011954	Dimas Prasetyo	43.38	0.9854	0.0069	0.1806	0.00045	FAIL*

Table 4. Watermark verification results across six attack conditions

Test Case	Sender / Tool	Hash Match	Robust BER	Robust NC	Nominal Extracted	Fragile Pass	Verdict
Original file	—	YES	0.000	1.000	Rp 5,950,000	Pass	Valid
WhatsApp (JPEG)	WhatsApp	NO	0.000	1.000	Rp 5,950,000	Pass	Compression
AI edit – Gemini	Gemini	NO	0.465	0.069	Rp 0 (lost)	Fail	Invalid
AI edit – GPT	ChatGPT	NO	0.472	0.056	Rp 0 (lost)	Fail	Invalid
AI edit – Grok	Grok	NO	0.375	0.250	Rp 0 (lost)	Fail	Invalid
+10% Python edit	Python script	NO	0.000	1.000	Rp 6,545,000	Fail	Invalid
+10 % Photoshop Edit (Baseline)	Adobe Photoshop	NO	0.000	1.000	Rp 6,545,000	Fail	Invalid

Based on Table 4, a single watermarked receipt (STR-20260429011606) was evaluated under six scenarios to assess the system's capability to distinguish authentic, recompressed, AI-manipulated, and forged documents. The results reveal distinct responses between conventional editing and generative AI tampering. Manual manipulation using Adobe Photoshop, which altered the transaction amount by +10%, successfully

triggered a "Fail" state in the fragile QIM-DCT layer but left the robust DWT-SVD watermark intact, maintaining BER = 0.000, NC = 1.000, and preserving the extracted nominal value of Rp 6,545,000. This indicates that traditional editing disrupts only localized integrity components without affecting the underlying watermark structure. Generative AI tools such as Gemini, ChatGPT, and Grok regenerated the entire image, overwriting the frequency-domain architecture that stores the robust watermark. This resulted in severe degradation of the DWT-SVD layer, with BER increasing to 0.375–0.472, NC dropping to as low as 0.056, and complete loss of the embedded nominal data (Rp 0 recovered). While benign recompression and manual forgery preserved the robust watermark (BER = 0.000), generative AI caused catastrophic signal destruction, making the collapse of the robust watermark an intrinsic and reliable indicator of AI-generated manipulation without requiring dedicated AI detection algorithms.

Ten synthetic emails were processed to evaluate the gateway's decision logic. Emails from registered senders with clean URLs were correctly allowed, whereas emails from unregistered senders were quarantined, confirming that sender registration is necessary but insufficient for delivery. Six emails were blocked, including four originating from registered accounts because they contained malicious URLs, demonstrating that URL analysis takes precedence over sender trust and effectively prevents whitelist exploitation attacks. This finding supports the observation of Nisha and Bakari that Business Email Compromise (BEC) attacks can bypass conventional trust mechanisms and therefore require active content inspection, such as malicious URL scanning, to mitigate fraud [31]. Overall, the two proposed modules provide a defense-in-depth architecture, where the email gateway intercepts threats before delivery, while the watermarking module authenticates documents after delivery and reliably detects manipulations, including those generated by AI tools.

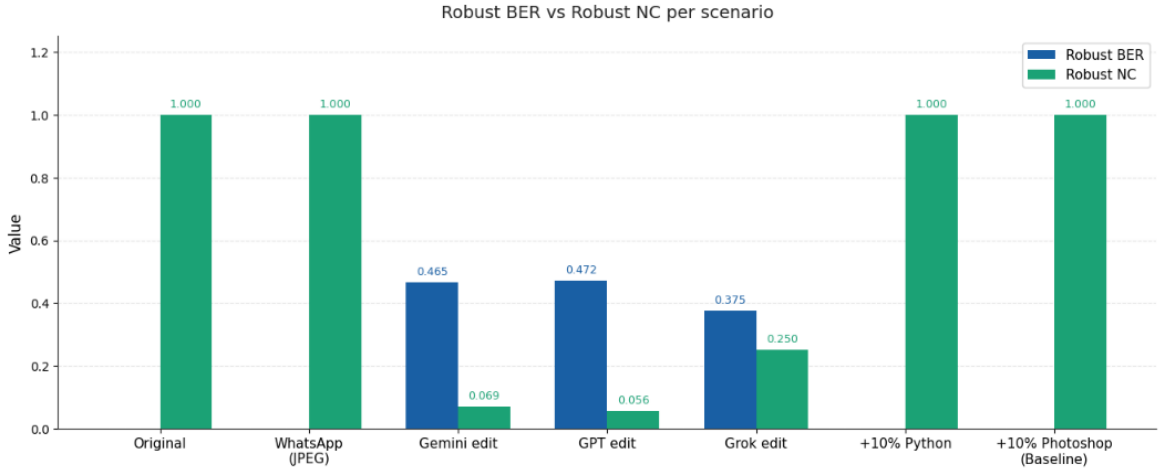

Figure 5. Visual comparison of the receipt under six verification scenarios. From left to right: original _wm.png, WhatsApp-recompressed JPEG, Gemini-edited, GPT-edited, Grok-edited, and Python +10% pixel manipulation

Table 5. Email gateway classification results across ten test scenarios

Sender	Registered	URLs Found	Malicious URLs	Receipt	Decision	Notes
budi@majubersama.co.id	YES	2	0	YES	PASS	Clean + registered
siti@karyamandiri.id	YES	2	0	NO	PASS	Clean + registered
unknown.buyer@gmail.com	NO	1	0	NO	QUARANTINE	Unknown sender
random.person@hotmail.com	NO	0	0	NO	QUARANTINE	Unknown, no URLs
ahmad.fauzi@kemenkeu.go.id	YES	2	1	NO	BLOCK	Malicious URL
billing-alert@fake-payment-id.xyz	NO	1	1	NO	BLOCK	Unknown + malicious
security-noreply@fake-bank-id.xyz	NO	3	3	NO	BLOCK	Full phishing
reza@bankmandiri.co.id	YES	2	1	NO	BLOCK	Registered, still blocked
dewi@ui.ac.id	YES	2	2	NO	BLOCK	GSB API v4 test
budi@majubersama.co.id	YES	3	1	NO	BLOCK	Mixed URLs

Although the initial evaluation in Table 5 employed a synthetic dataset of 10 emails, this controlled setup was necessary to validate the gateway's priority-based decision logic and its integration with the internal SQLite customer registry. The synthetic scenarios specifically tested Sender Verification (Step B) and demonstrated that malicious URL detection overrides sender trust. Emails 05, 08, 09, and 10 simulated compromised customer accounts and were successfully blocked despite originating from registered senders, confirming protection against whitelist exploitation. To assess performance under broader threats, the gateway was further evaluated using 50 phishing and 50 legitimate samples from public phishing email and malicious URL datasets (Table 6). The system achieved perfect Precision (1.000) for both email domain and URL

phishing detection, indicating zero false positives (FP = 0) and ensuring that legitimate business communications were not disrupted. However, the strict zero-trust configuration resulted in lower Recall values (0.41 for email domains and 0.52 for URLs), producing 20 false negatives and 12 false negatives, respectively. This trade-off stems from the VirusTotal-based detection mechanism, as newly emerging phishing domains and malicious URLs often lack sufficient consensus among security vendors at the time of scanning. Consequently, the gateway prioritizes absolute precision for authorized traffic over aggressive preemptive blocking, accepting lower recall to minimize disruption to normal business operations.

Table 6. Email gateway test using public dataset

Metric	Email domain detection (Total Sample 50)	URL phishing detection (Total Sample 50)
True positives (TP)	30	38
False positives (FP)	0	0
True negatives (TN)	50	50
False negatives (FN)	20	12
	Performance metrics	
Accuracy	0,555555556	0,611111111
Precision	1.000	1.000
Recall	0,416666667	0,527777778
F1 score	0,520833333	0,6

4. CONCLUSION

This study successfully designed, implemented, and evaluated a unified Dual-Shield framework that simultaneously addresses document fraud and communication-channel threats within a lightweight architecture. The proposed dual-layer watermarking engine achieved high imperceptibility (PSNR 43.38–43.48 dB; SSIM > 0.98) and maintained JPEG robustness across Q70–Q90 with BER < 0.01. Generative AI manipulations generated by Gemini, ChatGPT, and Grok were effectively detected without dedicated AI classifiers, as indicated by severe degradation of the robust watermark (BER = 0.375–0.472; NC as low as 0.056) caused by full-canvas image regeneration. In contrast, a targeted +10% numerical forgery triggered the semi-fragile QIM-DCT layer while preserving the robust DWT-SVD watermark (BER = 0.000), demonstrating complementary authentication capabilities. The email gateway correctly handled all ten synthetic scenarios and achieved perfect precision (1.000) for both email-domain and URL phishing detection, preventing whitelist exploitation by registered senders. This work contributes in four major ways. First, it introduces a dual-layer watermarking architecture tailored to transaction receipts by integrating a robust DWT-SVD identity layer with a semi-fragile QIM-DCT integrity layer, enabling differentiated verdicts (valid, compressed, and invalid) that cannot be achieved by either method alone. Second, it provides empirical evidence that conventional DWT-SVD degradation metrics can serve as a classifier-free indicator of generative AI document manipulation, addressing a gap in AI-based fraud detection research. Third, it proposes a pre-delivery email gateway whose priority-based logic allows malicious URLs to override sender trust, mitigating attacks not addressed by DMARC, SPF, and DKIM. Fourth, it presents a unified real-time framework that jointly secures communication channels and document authentication for SME and corporate environments. Several limitations remain. In the watermarking module, T5 failed because the embedding strength parameter (α) was optimized for JPEG robustness and imperceptibility, reducing resilience against salt-and-pepper noise, whereas T6 failed because DWT-SVD is inherently grid-aligned and lacks geometric invariance; addressing this would require synchronization mechanisms incompatible with the lightweight design objective. In the email gateway, the evaluation using only ten synthetic scenarios limits generalizability to more diverse phishing patterns. Moreover, the public-dataset evaluation yielded moderate recall (0.42 for email-domain detection and 0.53 for URL phishing) due to VirusTotal API detection latency, where newly emerging phishing domains and malicious URLs may lack sufficient vendor consensus. This precision–recall trade-off reflects the inherent limitation of third-party aggregation-based detection rather than a flaw in the gateway logic. Future work should enhance robustness against geometric attacks through synchronization-based watermarking and incorporate complementary mechanisms, such as heuristic domain-age analysis and autonomous URL sandboxing, to reduce false negatives while maintaining the zero-false-positive requirement.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Bonifacius Vicky Indriyono: Conceptualization, Methodology, Software, Writting – original draft Project administration, Funding. **Rabei Raad Ali**: Methodology, Review, Editing, Validation**,** Supervision.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no conflict interest.

DATA AVAILABILITY

Data will be made available on request.

REFERENCES

[1] F. Carroll, J. A. Adejobi, and R. Montasari, "How Good Are We at Detecting a Phishing Attack? Investigating the Evolving Phishing Attack Email and Why It Continues to Successfully Deceive Society," *SN Comput. Sci.*, vol. 3, no. 2, Mar. 2022, doi: 10.1007/s42979-022-01069-1.

[2] FBI, "Federal Bureau Of Investigation Crime Report," Dec. 2025.

[3] N. Gupta, "Security Risks of Generative AI in Financial Systems: A comprehensive review," *World Journal of Information Systems*, vol. 1, no. 3, pp. 17–24, Mar. 2025, doi: 10.17013/wjis.v1i3.16.

[4] E. Choo, M. Nabeel, R. De Silva, T. Yu, and I. Khalil, "A Large Scale Study and Classification of VirusTotal Reports on Phishing and Malware URLs," May 2022, [Online]. Available: http://arxiv.org/abs/2205.13155

[5] Ika Kurnia Sofiani, Mardiana, Nurlidia Putri, and Fitri Barokah, "CYBER RISK MANAGEMENT IN THE DIGITAL ERA: AN ANALYSIS OF MITIGATION STRATEGIES AND PREVENTIVE INNOVATIONS AGAINST CYBERCRIME IN INDONESIA," *Multidisciplinary Indonesian Center Journal (MICJO)*, vol. 2, no. 3, pp. 2600–2609, Jul. 2025, doi: 10.62567/micjo.v2i3.869.

[6] C. Diandra Athaillah, D. Larasati Tupen, A. Julianti Sari, F. Chalisha Adzzahra, and J. Indrawan, "TANTANGAN CYBER DEFENSE DALAM MENGHADAPI SERANGAN SIBER TERHADAP INFRASTRUKTUR PEMERINTAH DI INDONESIA," *GOVERNANCE: Jurnal Ilmiah Kajian Politik Lokal dan Pembangunan*, vol. 13, pp. 1–6, Feb. 2026.

[7] Z. Jiang, M. Guo, Y. Hu, J. Jia, and N. Z. Gong, "Certifiably Robust Image Watermark," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.04086

[8] S. Asli, A. Senol, E. Elbasi, A. E. Topcu, and N. Mostafa, "A Semi-Fragile, Inner-Outer Block-Based Watermarking Method Using Scrambling and Frequency Domain Algorithms," 2023, doi: 10.3390/electronics.

[9] P. V. Sanivarapu, K. N. V. P. S. Rajesh, K. M. Hosny, and M. M. Fouda, "Digital Watermarking System for Copyright Protection and Authentication of Images Using Cryptographic Techniques," *Applied Sciences (Switzerland)*, vol. 12, no. 17, Sep. 2022, doi: 10.3390/app12178724.

[10] S. E. Naffouti, A. Kricha, and A. Sakly, "A sophisticated and provably grayscale image watermarking system using DWT-SVD domain," *Visual Computer*, vol. 39, no. 9, pp. 4227–4247, Sep. 2023, doi: 10.1007/s00371-022-02587-y.

[11] Y. Okamoto, O. Genki, I. Yahiro, R. Hasegawa, P. Zhu, and H. Kataoka, "Image Generation and Learning Strategy for Deep Document Forgery Detection," Nov. 2023, [Online]. Available: http://arxiv.org/abs/2311.03650

[12] X. Zhang and Q. Su, "A spatial domain-based color image blind watermarking scheme integrating multilevel discrete Hartley transform," *International Journal of Intelligent Systems*, vol. 36, no. 8, pp. 4321–4345, Aug. 2021, doi: 10.1002/int.22461.

[13] A. Papathanasiou, G. Liontos, G. Paparis, V. Liagkou, and E. Glavas, "BEC Defender: QR Code-Based Methodology for Prevention of Business Email Compromise (BEC) Attacks," *Sensors*, vol. 24, no. 5, Mar. 2024, doi: 10.3390/s24051676.

[14] A. F. Naswir, L. Q. Zakaria, and S. Saad, "THE EFFECTIVENESS OF URL FEATURES ON PHISHING EMAILS CLASSIFICATION USING MACHINE LEARNING APPROACH," *Asia-Pacific Journal of Information Technology and Multimedia*, vol. 11, no. 2, pp. 49–58, Dec. 2022, doi: 10.17576/apjitm-2022-1102-04.

[15] P. V. Sanivarapu, K. N. V. P. S. Rajesh, K. M. Hosny, and M. M. Fouda, "Digital Watermarking System for Copyright Protection and Authentication of Images Using Cryptographic Techniques," *Applied Sciences (Switzerland)*, vol. 12, no. 17, Sep. 2022, doi: 10.3390/app12178724.

[16] S. E. Naffouti, A. Kricha, and A. Sakly, "A sophisticated and provably grayscale image watermarking system using DWT-SVD domain," *Visual Computer*, vol. 39, no. 9, pp. 4227–4247, Sep. 2023, doi: 10.1007/s00371-022-02587-y.

[17] F. Anwar, A. Fadlil, and D. I. Riadi, "ANALISIS VALIDASI FILE UPLOAD MENGGUNAKAN METADATA PNG PADA APLIKASI BERBASIS WEB," *Jurnal Informatika dan Komputer)*, vol. 6, no. 2, pp. 185–193, 2022.

[18] E. Choo, M. Nabeel, R. De Silva, T. Yu, and I. Khalil, "A Large Scale Study and Classification of VirusTotal Reports on Phishing and Malware URLs," May 2022, [Online]. Available: http://arxiv.org/abs/2205.13155

[19] X. Zhou, H. Zhang, and C. Wang, "A robust image watermarking technique based on DWT, APDCBT, and SVD," *Symmetry (Basel).*, vol. 10, no. 3, Mar. 2018, doi: 10.3390/SYM10030077.

[20] Y. Zhang, C. Wang, and X. Zhou, "RST resilient watermarking scheme based on DWT-SVD and scale-invariant feature transform," *Algorithms*, vol. 10, no. 2, Jun. 2017, doi: 10.3390/a10020041.

[21] M. Begum, J. Ferdush, and M. S. Uddin, "A Hybrid robust watermarking system based on discrete cosine transform, discrete wavelet transform, and singular value decomposition," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 5856–5867, Sep. 2022, doi: 10.1016/j.jksuci.2021.07.012.

[22] M. Rahardi, A. Aminuddin, F. F. Abdulloh, N. Sholihah, and F. Ernawan, "Optimizing Imperceptibility and Robustness Using Selective DCT Coefficients for Image Copyright Protection," *Engineering, Technology and Applied Science Research*, vol. 15, no. 4, pp. 24825–24831, Aug. 2025, doi: 10.48084/etasr.10241.

[23] D. L. Donoho and X. Huo, "Uncertainty Principles and Ideal Atomic Decomposition," 2001.

[24] J. Zhang, J. Du, X. Xi, and Z. Yang, "Chebyshev Chaotic Mapping and DWT-SVD-Based Dual Watermarking Scheme for Copyright and Integrity Authentication of Remote Sensing Images," *Symmetry (Basel).*, vol. 16, no. 8, Aug. 2024, doi: 10.3390/sym16080969.

[25] X. Liu, R. Xu, and Y. Chen, "Securing Digital Media Integrity: A Survey of Watermarking and Manipulation Detection for Image Authentication," Oct. 30, 2025. doi: 10.36227/techrxiv.176186513.32946978/v1.

[26] I. Assini, A. Badri, K. S. A. Sahel, and A. Baghdad, "A robust hybrid watermarking technique for securing medical image," *International Journal of Intelligent Engineering and Systems*, vol. 11, no. 3, pp. 169–176, 2018, doi: 10.22266/IJIES2018.0630.18.

[27] M. Jin and L. Shi, "Research of moving targets detection and identification," in *2009 2nd International Conference on Intelligent Computing Technology and Automation, ICICTA 2009*, 2009, pp. 332–335. doi: 10.1109/ICICTA.2009.316.