Penelusuran Asal dan Deteksi Manipulasi Citra pada Kanal Whatsapp Menggunakan *Watermark* Ganda DWT dan Tanda Tangan Pascakuantum Ml-DSA-44

Alif Maulana Arifin[1*], Dedy Suhariyanto[1], Ferdiansyah Winardi[1]

[1]*Program Studi Teknik Informatika S-2, Program Pascasarjana, Universitas Pamulang, Tangerang Selatan, Indonesia*
Email korespondensi: alifmaulana26@gmail.com[*]

Received: 2026-10-20 | Accepted: 2026-10-27 | Published: 2026-10-28

Abstrak

Citra yang dikirim melalui WhatsApp kehilangan metadata dan dikompresi ulang, sehingga *hash* atau berkas tanda tangan pendamping tidak lagi dapat dipakai untuk membuktikan asal dan keutuhannya. Artikel ini mengusulkan skema *watermark* ganda pada domain *Discrete Wavelet Transform* (*DWT*): pengenal *robust* 32 bit yang menelusuri sertifikat penerbit, dan *fingerprint* 256 bit *semi-fragile* per blok 32×32 piksel yang dibandingkan lewat *Bit Error Rate* (*BER*) untuk melokalisasi manipulasi secara *blind*. Sertifikat ditandatangani dengan tanda tangan pascakuantum ML-DSA-44 (FIPS 204) dan disimpan di server, sedangkan piksel hanya membawa pengenalnya. Pada pengiriman nyata melalui WhatsApp Web terhadap 60 citra, pengenal terbaca dan tanda tangan valid pada seluruhnya, tidak ada citra utuh yang dinilai *malicious* (0/48) sedangkan pencocokan *hash* menolak seluruhnya, dan manipulasi lokal terdeteksi pada 9 dari 12 citra. Pada aplikasi seluler, kompresi lebih berat (*BER* sekitar 0,28): pengenal tetap terbaca pada 10 dari 10 citra, tidak ada citra utuh yang dinilai malicious (0/7), tetapi tiga citra termanipulasi juga tidak terdeteksi (0/3). Dengan demikian asal citra dapat dibuktikan di WhatsApp, sedangkan deteksi manipulasi baru terukur pada kompresi ringan dan memerlukan kalibrasi ambang per kanal. Pada 200 citra *Chest X-Ray* tanpa kanal, presisi lokalisasi 1,0000 dan *recall* 0,99995 dengan *PSNR* 32,9 dB. Ablasi pada 48 citra tanpa kompresi menunjukkan bahwa BER dasar citra utuh berasal dari pembulatan 8-bit dan pemotongan (clipping) piksel di area sangat gelap atau sangat terang, bukan dari fingerprint itu sendiri.

Kata kunci: *Autentikasi Citra, Discrete Wavelet Transform, ML-DSA, Watermark Ganda, WhatsApp.*

Abstract

Images sent through WhatsApp lose their metadata and are recompressed, so a hash or a companion signature file can no longer prove their origin and integrity. This article proposes a dual-watermark scheme in the Discrete Wavelet Transform (DWT) domain: a 32-bit robust identifier that traces back to the issuer's certificate, and a semi-fragile 256-bit fingerprint per 32×32-pixel block compared through Bit Error Rate (BER) to localize manipulation blindly. The certificate is signed with the post-quantum signature ML-DSA-44 (FIPS 204) and stored on a server, while the pixels carry only the identifier. In a real WhatsApp Web transfer of 60 images, the identifier is read and the signature is valid for all of them, no intact image is judged malicious (0/48) while hash matching rejects all of them, and local manipulation is detected in 9 of 12 images. On the mobile app, compression is heavier (BER about 0.28): the identifier is still read in 10 of 10 images, no intact image is judged malicious (0/7), but the three manipulated images are not detected either (0/3). Hence image origin can be proven over WhatsApp, whereas manipulation detection has only been measured under light compression and needs per-channel threshold calibration. *On 200 Chest X-Ray images without a channel, localization precision is 1.0000 and recall 0.99995 at a PSNR of 32.9 dB.* An ablation on 48 uncompressed images shows that the baseline BER of intact images comes from 8-bit rounding and pixel clipping in very dark or very bright areas, not from the fingerprint itself.

Keywords: *Dual Watermark, Image Authentication, WhatsApp, Discrete Wavelet Transform, ML-DSA*

1. Pendahuluan

WhatsApp dan aplikasi perpesanan serupa lazim dipakai untuk mengirim citra, termasuk hasil pemeriksaan medis dan dokumen identitas, melalui kanal yang mengkompresi ulang citra dan membuang metadata [1], [2], [3]. Pada saat yang sama, citra semakin mudah dimanipulasi dengan perangkat penyunting maupun model generatif [4], [5]. Ketika citra berfungsi sebagai bukti diagnosis atau bukti

forensik, penerima perlu dapat memastikan dua hal: siapa penerbit citra itu, dan apakah isinya berubah sejak diterbitkan.

Pendekatan klasik tidak cocok untuk kanal ini. Skema *fragile* menyisipkan *hash* kriptografis ke dalam citra dan menolak seluruh citra bila satu bit berbeda [6], [7]. Karena SHA-256 memiliki *avalanche effect*, perubahan pembulatan akibat kompresi ulang yang sah tidak dapat dibedakan dari manipulasi yang disengaja. Tanda tangan digital pada berkas, atau berkas sertifikat pendamping, hilang atau menjadi tidak cocok begitu WhatsApp membuang metadata dan menyimpan ulang citra sebagai JPEG [8], [3]. Skema *semi-fragile* mentoleransi distorsi jinak [9] dan menyisipkan informasi pada domain transformasi seperti *Discrete Cosine Transform* (*DCT*) dan *Discrete Wavelet Transform* (*DWT*) [10], tetapi umumnya belum menjawab pertanyaan penerbit: siapa yang mengesahkan citra ini dan dapatkah itu dibuktikan kepada pihak ketiga setelah citra melewati kanal.

Penelitian terdekat adalah skema *DCT semi-fragile* untuk citra medis oleh Hacini dkk. [11], yang memakai SHA3-256 dan AES-256-CTR serta pengklasifikasi *Random Forest* tiga kelas (*intact*, *benign*, *malicious*), dengan *PSNR* 44,2 dB dan *SSIM* di atas 0,98. Skema tersebut sepenuhnya simetris sehingga tidak menyediakan *non-repudiation*, dan bersifat *non-blind* karena membutuhkan citra ber-*watermark* asli sebagai referensi, keterbatasan yang diakui sendiri oleh penulisnya. Karya lain mengisi sebagian aspek: kerangka *dual-watermark* ringan [12] belum melokalisasi *tampering*, skema *semi-fragile* untuk citra warna [13] dan *block mapping* dengan *self-recovery* [14] tidak membahas tanda tangan asimetris, dan skema spasial *blind* [15] melaporkan uji lewat WhatsApp tetapi bersifat *robust*, bukan *semi-fragile*. Tanda tangan pascakuantum ringan telah diusulkan untuk perangkat IoT [16] dan distandarkan NIST sebagai *ML-DSA* [17], tetapi pada sumber yang kami tinjau belum ada skema watermark citra berlapis ganda yang terhubung ke tanda tangan pascakuantum dan diuji pada kompresi aplikasi pesan instan. Yang ada baru tinjauan umum watermarking tahan kuantum [20] dan karya yang memadukan tanda tangan pascakuantum dengan media pada tataran kerangka kerja atau kinerja kriptografi: kerangka pencegahan *deepfake* dengan *watermarking* hibrida dan blockchain [25], perbandingan kinerja tanda tangan pascakuantum untuk media [26], dan autentikasi citra medis dengan tanda tangan pascakuantum [27] (strategi pencarian pada Subbab K). Deteksi *copy-move* pasif [18] tidak memberi bukti penerbit karena tidak menyisipkan apa pun sejak penerbitan. Tabel 1 merangkum posisi penelitian ini.

Tabel 1. Posisi penelitian ini terhadap penelitian terdahulu.

Aspek	Hacini	Taj	Basha	Penelitian ini
Citra medis	Ya	Ya	Tidak	Ya
Toleran kompresi sah (*semi-fragile*)	Ya	Ya	Tidak (*robust*)	Ya
Lokalisasi tamper per blok	Ya	Rencana	Tidak	Ya
Verifikasi tanpa citra asli (blind)	Tidak	Ya	Ya	Ya
Tanda tangan digital asimetris	Tidak	Tidak	Tidak	Ya
Kriptografi pasca-kuantum	Sebagian (simetris)	Tidak	Tidak	Ya (ML-DSA-44)
Anti-splicing berbasis posisi blok	Tidak	Tidak	Tidak	Ya
Uji kanal WhatsApp nyata	Tidak	Tidak	Ya	Ya (70 citra)

Artikel ini mengusulkan skema yang membuat asal dan keutuhan citra dapat diperiksa dari satu berkas tunggal setelah melewati WhatsApp. Kontribusinya adalah: (1) skema *watermark* ganda pada domain *DWT*, yaitu pengenal *robust* yang menelusuri sertifikat bertanda tangan ML-DSA-44 di server dan *fingerprint semi-fragile* per blok yang melokalisasi manipulasi tanpa citra asli; (2) evaluasi pada kanal nyata, yaitu pengiriman WhatsApp Web terhadap 60 citra dan uji pembanding pada aplikasi seluler, dibandingkan dengan alur *hash*; (3) temuan bahwa putusan integritas bergantung pada kanal dan bahwa derau tingkat blok menumpuk di area datar, sementara penelusuran asal tetap andal; dan (4) implementasi referensi mikroservis yang menunjukkan skema ini dapat dijalankan, dengan kunci privat terisolasi pada satu komponen. Tanda tangan disimpan di server, bukan di piksel, sehingga pilihan ML-DSA-44 tidak memengaruhi hasil eksperimen pada piksel; peran dan batasannya dibahas pada Subbab J. Bagian berikutnya menjelaskan skema dan protokol uji, hasil dan pembahasan, serta kesimpulan.

2. Metode Penelitian *(Methodology / Research Method)*

A. Pendekatan penelitian

Penelitian ini bersifat eksperimental: skema dirancang, diimplementasikan, lalu diuji pada simulasi kompresi dan pada kanal WhatsApp nyata. Cakupan dibatasi pada pembuktian integritas berkas dan otentisitas penerbit, bukan kebenaran fisik objek yang dicitrakan. Transformasi geometris agresif seperti rotasi, *cropping*, dan *resize* berada di luar cakupan karena merusak korelasi *position-salt*.

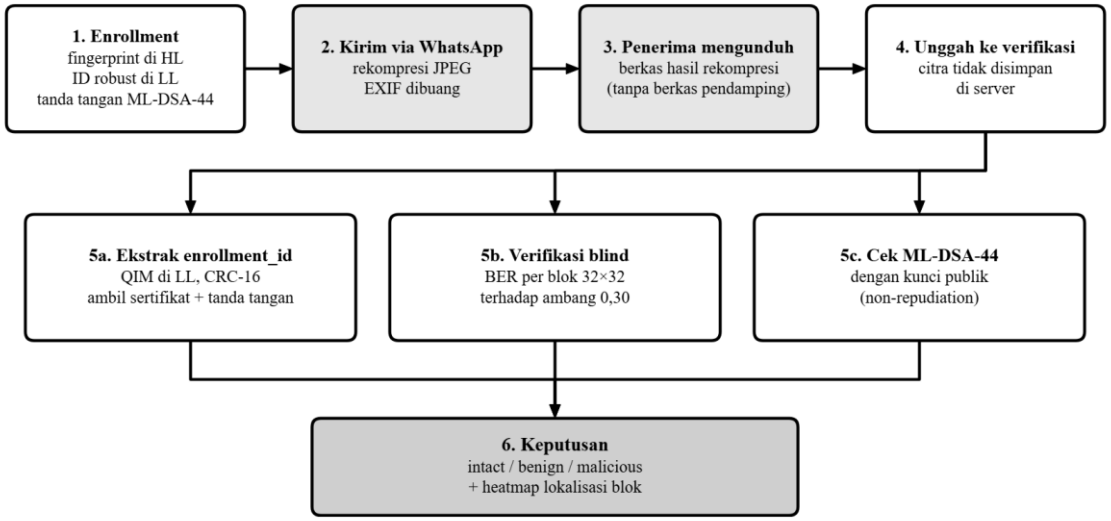
B. Landasan teori ringkas

Discrete Wavelet Transform (*DWT*) mendekomposisi citra menjadi sub-pita aproksimasi frekuensi rendah (*LL*) dan tiga sub-pita detail (*LH*, *HL*, *HH*) yang terlokalisasi secara spasial dan frekuensial [10]. Koefisien *LL* relatif stabil terhadap kompresi ringan sehingga cocok sebagai dasar *fingerprint*, sedangkan koefisien detail menjadi tempat penyisipan yang kurang tampak oleh pengamat. Penyisipan bit memakai *Quantization Index Modulation* (*QIM*): setiap bit menentukan kisi kuantisasi (genap atau ganjil) tempat koefisien dibulatkan, dan langkah kuantisasi yang lebih besar membuat bit lebih tahan terhadap kompresi dengan imbalan distorsi yang lebih besar.

Watermark fragile dirancang agar rusak oleh perubahan sekecil apa pun, *watermark robust* agar bertahan dari distorsi berat, dan *semi-fragile* berada di antaranya: toleran terhadap distorsi jinak tetapi sensitif terhadap manipulasi bermakna [9]. *Bit Error Rate* (*BER*), yaitu proporsi bit *fingerprint* yang berbeda, memberi ukuran gradual yang tidak dimiliki *hash* kriptografis. ML-DSA (FIPS 204) adalah standar tanda tangan digital berbasis kisi yang tahan terhadap serangan komputer kuantum dan menyediakan *non-repudiation* melalui pasangan kunci asimetris [17]. Pada kanal WhatsApp, citra mode standar dikompresi ulang sebagai JPEG dengan metadata EXIF dibuang, dan tingkat kompresinya berbeda antarperangkat [8], [3]. Fingerprint pada skema ini bukan keluaran hash, melainkan paritas koefisien LL yang dikuantisasi (Subbab D), sehingga gradasi BER terjaga.

C. Skema usulan

Skema memisahkan apa yang dibawa piksel dari apa yang disimpan server (Gambar 1). Piksel membawa dua lapis *watermark*: lapisan *robust* berisi *enrollment_id* 32 bit yang tahan kompresi berat, dan lapisan *semi-fragile* berisi *fingerprint* per blok untuk lokalisasi. Server menyimpan metadata *enrollment*, yaitu *sertifikat* SHA-256 per blok, *root hash*, dan tanda tangan ML-DSA-44 atas *root hash*. Pembagian ini dipaksa oleh ukuran: tanda tangan ML-DSA-44 berukuran 2.420 byte dan kunci publiknya 1.312 byte [17], terlalu besar untuk disisipkan ke piksel tanpa merusak citra. Dengan pengenal pendek, penerima yang hanya memegang satu berkas gambar tetap dapat menelusuri tanda tangan penerbit.



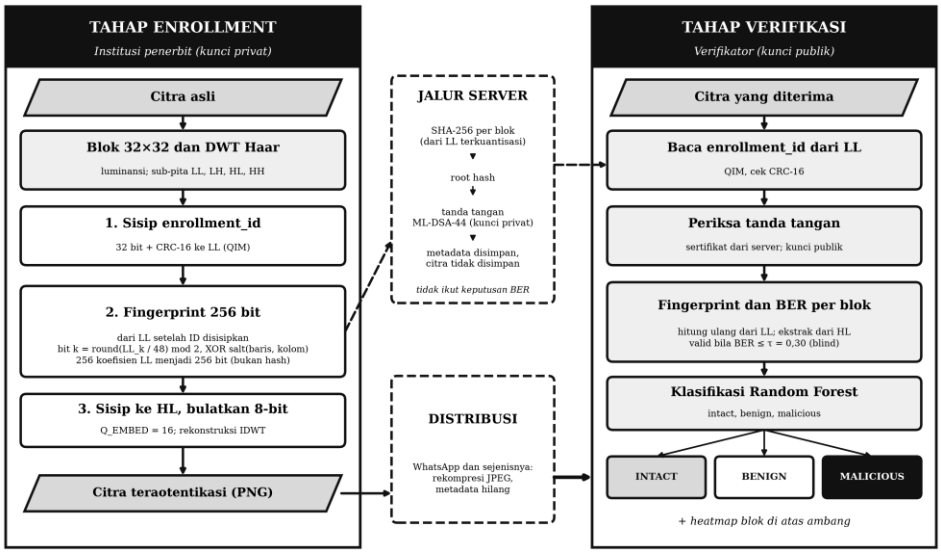
Gambar 1. Skema usulan dan alur pengiriman melalui WhatsApp

Alur WhatsApp berjalan sebagai berikut. Penerbit melakukan *enrollment* dan menyimpan keluaran dalam format *lossless* (PNG). Citra dikirim lewat WhatsApp pada mode citra standar, bukan sebagai dokumen. Penerima mengunduh berkas hasil kompresi ulang lalu mengunggahnya ke verifikasi. Sistem mengekstrak *enrollment_id* dari piksel, mengambil metadata, menghitung *BER* per blok, memeriksa tanda tangan, dan menghasilkan status serta *heatmap* lokalisasi. Verifikasi tidak memerlukan citra asli maupun berkas pendamping.

D. Algoritma enrollment dan verifikasi

Pada *enrollment*, citra dipisahkan menjadi kanal luminansi dan kanal warna, lalu luminansi dipotong menjadi blok 32×32 piksel. Setiap blok didekomposisi dengan DWT Haar satu level menjadi sub-pita LL, LH, HL, dan HH, sehingga LL berisi 16×16 = 256 koefisien. Lapisan robust (*enrollment_id*) disisipkan

lebih dulu ke LL (Subbab E). *Fingerprint* 256 bit kemudian dihitung dari LL yang sudah memuat pengenal itu: bit ke-k adalah paritas koefisien LL ke-k setelah dikuantisasi dengan langkah Q_LL_HASH, yaitu f(k) = [round(LL(k) / Q_LL_HASH) mod 2] ⊕ s(k), dengan s(k) adalah *position-salt* yang diturunkan dari koordinat baris dan kolom blok, sehingga blok yang disalin ke posisi lain di citra yang sama menghasilkan *fingerprint* yang tidak cocok. *Fingerprint* ini bukan keluaran fungsi hash: satu koefisien yang melewati batas kuantisasi hanya membalik satu dari 256 bit, sehingga *BER* berubah bertahap, sedangkan pada SHA-256 *avalanche effect* membalik sekitar separuh bit keluaran. Bit *fingerprint* disisipkan ke koefisien HL dengan kekuatan Q_EMBED, lalu blok direkonstruksi dan dibulatkan ke bilangan bulat 8-bit. SHA-256 hanya dipakai pada jalur server: hash tiap blok dihimpun menjadi *root hash* yang ditandatangani dengan ML-DSA-44, dan tidak ikut dalam keputusan *BER*. Alur lengkap ditunjukkan pada Gambar 2 dan ringkasan langkah pada Tabel 2.



Gambar 2. Alur *enrollment* dan verifikasi. Jalur piksel (*fingerprint* dari paritas kuantisasi LL, bukan hash) dipisahkan dari jalur server (SHA-256 dan ML-DSA-44)

Pada verifikasi, *fingerprint* dihitung ulang langsung dari *LL* citra yang diuji dan dibandingkan dengan bit yang diekstraksi dari *HL*. Verifikasi bersifat blind: tidak ada citra asli yang perlu disimpan. Untuk blok ke-*b*, *BER* dihitung pada Persamaan (1) dan blok dinyatakan valid bila memenuhi Persamaan (2) dengan ambang $\tau = 0{,}30$.

$$BER_b = (1/256)\ \Sigma\ [\ f_b(i)\ \oplus f'_b(i)\ ],\ \ i = 1, \ldots, 256 \quad (1)$$

$$valid_b\ \Leftrightarrow\ BER_b \leq \tau \quad (2)$$

Ciri *BER* per blok diringkas menjadi lima fitur per citra untuk pengklasifikasi *Random Forest* tiga kelas (*intact*, *benign*, *malicious*), yang rancangan fiturnya diadaptasi dari Hacini dkk. [11]. Pemisahan peran ini sengaja: keputusan toleransi dan lokalisasi memakai *BER* pada *fingerprint* mentah, sedangkan SHA-256 dan ML-DSA-44 berperan sebagai bukti forensik dan *non-repudiation* yang independen dari keputusan piksel. Tanda tangan tetap valid pada citra yang pikselnya dimanipulasi, dan *sertifikat* yang ditandatangani kunci institusi lain terdeteksi tidak valid walau pikselnya utuh. Kedua perilaku ini diperiksa lewat uji unit *adversarial*. Pengklasifikasi dilatih dari 5.200 sampel yang dibangkitkan dari 200 citra sumber dan seluruhnya memuat lapisan *enrollment_id*, karena lapisan itu menggeser BER dasar citra utuh (bagian Hasil dan Pembahasan, Subbab B); model yang dilatih tanpa lapisan tersebut menilai sebagian besar citra utuh produksi sebagai benign.

Tabel 2. Ringkasan algoritma enrollment dan verifikasi

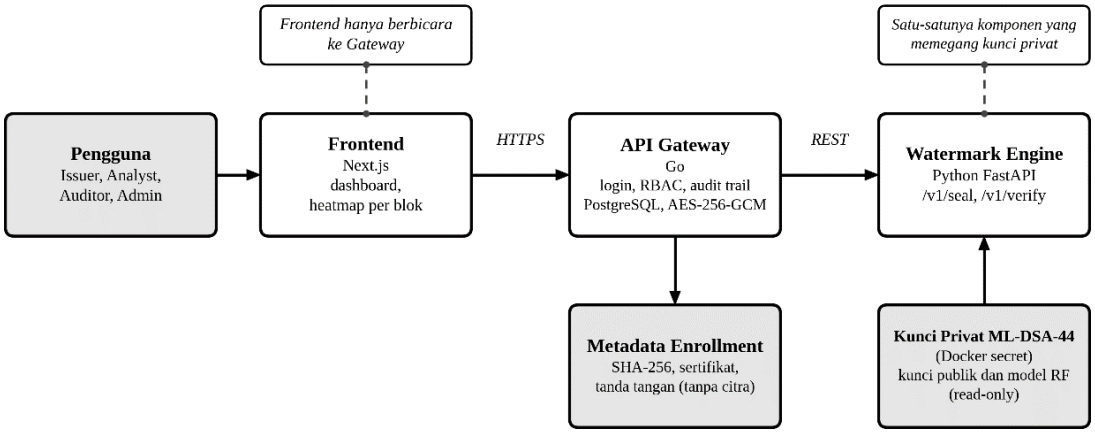
Langkah	Proses
Enrollment 1	Pisahkan luminansi dan kanal warna; potong luminansi menjadi blok 32×32 piksel
Enrollment 2	DWT Haar satu level pada tiap blok (LL, LH, HL, HH)
Enrollment 3	Sisipkan *enrollment_id* (32 bit + CRC-16) ke LL dengan QIM (Q_ID_LL)
Enrollment 4	*Fingerprint* 256 bit dari LL yang sudah memuat ID: round(LL / Q_LL_HASH) mod 2, XOR *position-salt* (baris, kolom)
Enrollment 5	Sisipkan bit *fingerprint* ke HL (Q_EMBED), rekonstruksi, bulatkan ke 8-bit
Enrollment 6	Hitung SHA-256 per blok; tandatangani *root hash* dengan ML-DSA-44; simpan metadata, jangan simpan citra
Verifikasi 1	Ekstrak enrollment_id; ambil sertifikat dan tanda tangan dari server
Verifikasi 2	Hitung ulang fingerprint dari LL citra uji dengan position-salt yang sama
Verifikasi 3	Ekstrak bit dari HL; hitung BER per blok (Persamaan 1); tandai blok dengan BER > τ
Verifikasi 4	Periksa tanda tangan ML-DSA-44 dengan kunci publik institusi
Verifikasi 5	Klasifikasikan citra (*intact, benign, malicious*) dan keluarkan heatmap blok

E. Lapisan *robust enrollment_id*

Lapisan *robust* membawa *enrollment_id* sepanjang 32 bit ditambah CRC-16. Pengenal disisipkan dengan QIM pada koefisien LL (langkah Q_ID_LL = 16, dither ID_DITHER = 4 agar lattice tidak berimpit dengan kuantisasi *fingerprint*), diulang lima salinan per blok, dan suara dari seluruh blok dijumlahkan saat ekstraksi. Verifikator membaca *enrollment_id* dari piksel lalu mengambil sertifikat dan tanda tangan dari server, sehingga verifikasi berjalan walau metadata berkas hilang. Pengenal yang pendek memungkinkan redundansi tinggi tanpa merusak kualitas visual. Karena lapisan ini disisipkan lebih dulu dan *fingerprint* dihitung dari LL yang sudah memuatnya, lapisan ini tidak membalik bit *fingerprint* saat *enrollment*; pengaruh tidak langsungnya pada BER dasar diukur pada Subbab B bagian Hasil dan Pembahasan.

F. Implementasi referensi

Skema diimplementasikan sebagai tiga komponen (Gambar 3). *Frontend* Next.js hanya berbicara ke API *Gateway* Go dan tidak pernah memanggil *engine* Python secara langsung. *Gateway* menangani autentikasi, otorisasi berbasis peran, *audit trail* pada PostgreSQL, dan penerusan permintaan ke *engine* melalui REST. *Engine* Python (FastAPI, OpenCV, PyWavelets, scikit-learn, *dilithium-py*) melakukan seluruh pemrosesan citra, penandatanganan, dan klasifikasi, dan dikemas sebagai kontainer Docker. Kunci privat ML-DSA-44 hanya dipasang sebagai *Docker secret* pada kontainer *engine*, sedangkan kunci publik dan model dipasang *read-only*. Jalur verifikasi hanya membutuhkan kunci publik, sehingga pemeriksa tidak pernah menyentuh kunci privat.

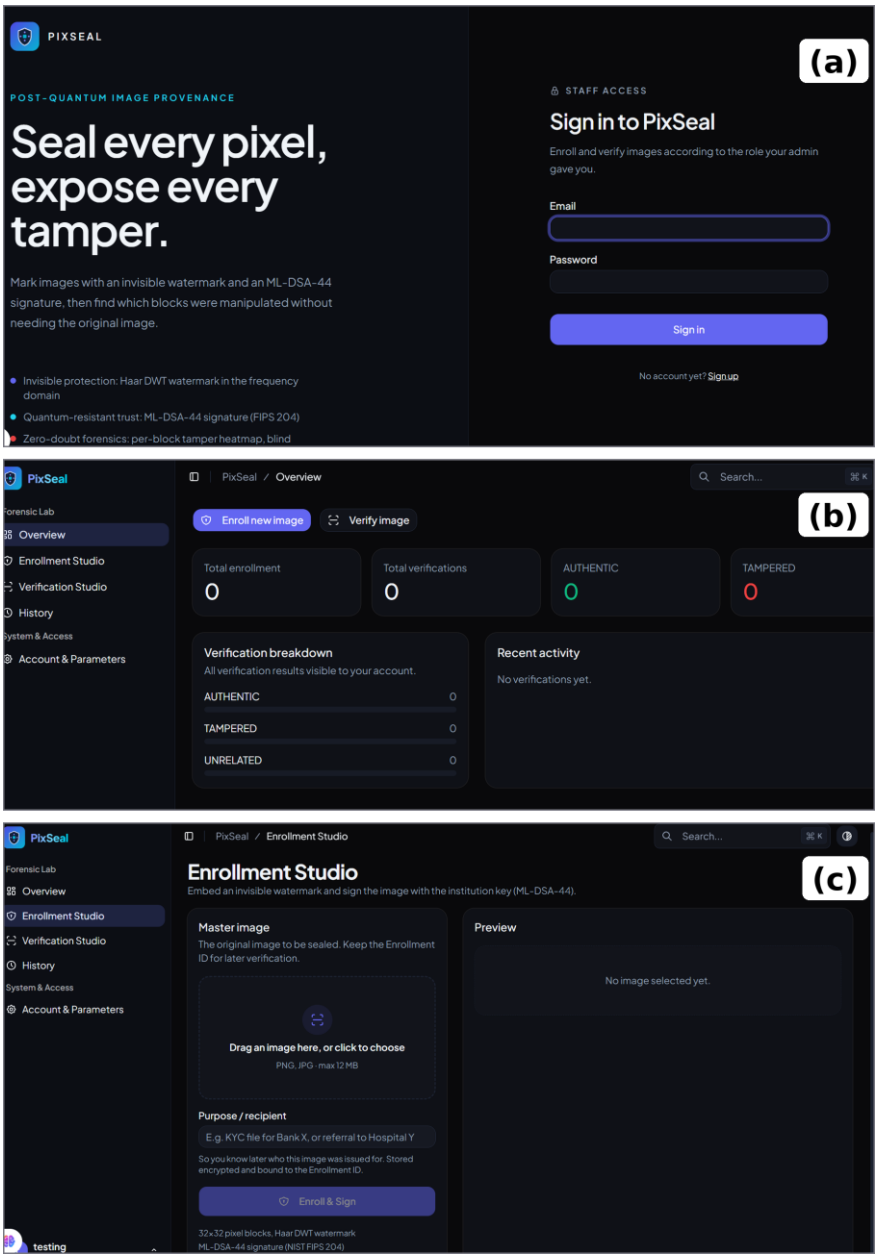


Gambar 3. Arsitektur implementasi referensi

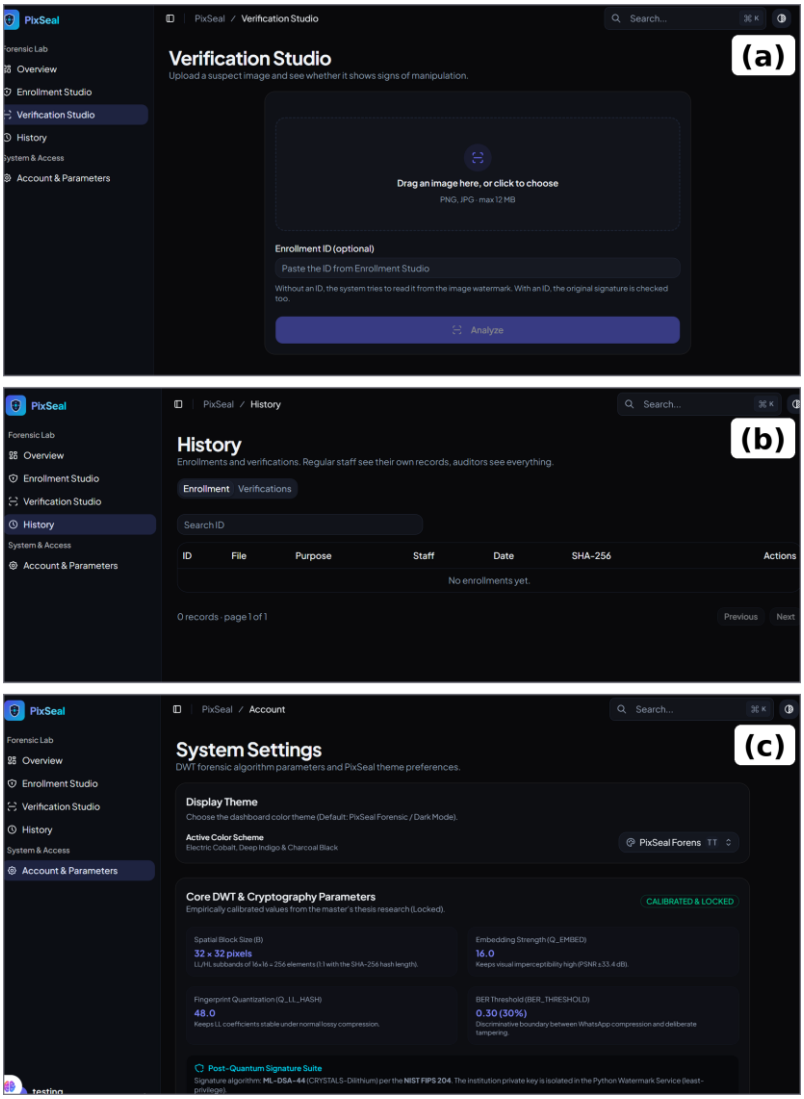
Untuk meminimalkan data pribadi, server tidak menyimpan citra. Basis data hanya memuat metadata *enrollment* (SHA-256, dimensi, *sertifikat*, tanda tangan). Citra ber-*watermark* ditahan sementara di memori (*TTL* 10 menit) agar penerbit dapat mengunduhnya sekali, nama berkas dan tujuan *enrollment* disegel dengan AES-256-GCM, dan citra yang diverifikasi tidak disimpan. Kontrak antarlayanan bersifat *stateless* per permintaan: *endpoint seal* mengembalikan citra ber-*watermark*, *sertifikat*, dan tanda tangan, serta *endpoint verify* mengembalikan status, probabilitas kelas, *BER* per blok, serta hasil pemeriksaan tanda tangan.

ISSN: 2252-4517
EISSN: 2723-7249

Antarmuka implementasi referensi ditunjukkan pada Gambar 4 dan Gambar 5, berupa tangkapan layar dari akun uji dengan data yang masih kosong. Halaman masuk dibatasi untuk staf (Gambar 4a). Halaman ringkasan menampilkan jumlah enrollment dan verifikasi beserta sebaran status AUTHENTIC, TAMPERED, dan UNRELATED (Gambar 4b). Enrollment Studio menerima citra master dan tujuan penerbitan, lalu menyisipkan watermark dan menandatangani citra dengan kunci institusi ML-DSA-44 (Gambar 4c). Verification Studio menerima citra yang dicurigai dengan Enrollment ID opsional: tanpa ID, sistem membaca pengenal dari watermark pada citra, dan dengan ID, tanda tangan asli ikut diperiksa (Gambar 5a). Halaman riwayat memisahkan catatan enrollment dan verifikasi; staf biasa hanya melihat catatannya sendiri, sedangkan auditor melihat seluruhnya (Gambar 5b). Halaman parameter menampilkan parameter DWT dan kriptografi yang sudah dikalibrasi dan dikunci (Gambar 5c).



Gambar 4. Antarmuka implementasi referensi (1): (a) halaman masuk staf; (b) halaman ringkasan; (c) Enrollment Studio



Gambar 5. Antarmuka implementasi referensi (2): (a) Verification Studio; (b) halaman riwayat; (c) halaman parameter sistem

G. Data dan protokol uji

Data sekunder berupa 200 citra *Chest X-Ray* publik [19] (lisensi CC BY 4.0, pengambilan sampel acak dengan *seed* 42), dipakai untuk lokalisasi dan klasifikasi tanpa kanal. Untuk uji kanal nyata disiapkan 60 citra PNG (sisi terpanjang 1.024 piksel) yang di-*enroll* pada sistem: 48 dikirim apa adanya dan 12 dimanipulasi lokal (tempel blok 64×64 piksel dari area lain) setelah *enrollment*. Citra dikirim lewat WhatsApp Web (peramban) sebagai foto dalam dua batch berurutan, lalu berkas hasil unduhan diverifikasi *(peramban,* sistem operasi, dan opsi kualitas foto: Google Chrome 155.0.8059.85 pada Windows 11, kualitas foto standar). Sebagai pembanding, 10 citra terpilih (7 utuh, 3 termanipulasi) dikirim lewat aplikasi WhatsApp seluler ke sebuah grup, lalu diunduh melalui WhatsApp Web *(*Infinix Note 50, Android 16, WhatsApp 2.26.41.77, kualitas foto standar). Sapuan kompresi JPEG simulasi (kualitas 40–90), *resize*, dan citra tak terkait dijalankan pada 6 citra melalui layanan yang sama, sedangkan kalibrasi lapisan *robust* memakai 9 *crop* berukuran 512×512 piksel.

Protokol uji kanal nyata berjalan dalam lima langkah. Pertama, 48 citra diproses lewat *enrollment* dan keluaran (PNG) disimpan tanpa kompresi ulang. Kedua, 12 dari citra tersebut disalin lalu dimanipulasi lokal sehingga menjadi 60 berkas. Ketiga, seluruh berkas dikirim lewat WhatsApp Web sebagai foto dalam dua batch berurutan dan hasil unduhannya disimpan. Keempat, setiap berkas diverifikasi tanpa memberi *enrollment_id*, sehingga pengenal harus dibaca dari piksel. Kelima, berkas yang sama dibandingkan dengan alur tanpa sistem (kecocokan SHA-256 terhadap berkas asli). Uji pembanding aplikasi seluler mengulang langkah kedua sampai kelima pada 10 citra terpilih, dengan pencocokan setiap berkas ke aslinya lewat

ukuran gambar dan kemiripan piksel karena urutan terima tidak dijamin. Pengiriman berulang (*forward*), tangkapan layar, dan foto ulang dari layar belum diuji.

H. Parameter, skenario, dan metrik

Parameter algoritma tercantum pada Tabel 3 dan skenario pengujian pada Tabel 4.

Tabel 3. Parameter algoritma inti dan lapisan robust

Parameter	Nilai	Keterangan
Wavelet	Haar	DWT satu level
Ukuran blok	32×32 piksel	Selaras dengan *fingerprint* 256 bit
Q_EMBED	16,0	Kekuatan sisip ke HL (kalibrasi empiris)
Q_LL_HASH	48,0	Kuantisasi LL untuk *fingerprint*
Ambang BER (τ)	0,30	Batas blok valid
Tanda tangan	ML-DSA-44	FIPS 204, NIST Level 1
enrollment_id	32 bit + CRC-16	Lapisan *robust*, 5 salinan per blok
Q_ID_LL / ID_DITHER	16,0 / 4,0	QIM pada LL (kalibrasi awal, 9 citra)
Random Forest	100 pohon, kedalaman maks. 10	Tiga kelas, *seed* 42

Tabel 4. Skenario pengujian

Kel.	Skenario	Ekspektasi
A	Citra hasil *enrollment* tanpa modifikasi	Tidak *malicious (intact atau benign),* tanda tangan valid
B	Kompresi JPEG simulasi (kualitas 40–90)	*Benign* hingga sekitar q60, *malicious* bertahap di bawahnya
C	Manipulasi piksel lokal (2–4 blok)	*Malicious*, blok teredit BER > 0,30
D	Citra lain tanpa *watermark*	Ditolak, BER mendekati 50%
E	Perubahan ukuran (*resize*)	Ditolak terkendali, batas desain
F	Pengiriman WhatsApp nyata (48 utuh, 12 dimanipulasi)	ID terbaca, utuh tidak *malicious*, manipulasi *malicious*
G	Foto ulang atau tangkapan layar	Ditolak terkendali, batas desain (belum diuji)

Penelusuran asal diukur dari proporsi berkas yang *enrollment_id*-nya terbaca dan tanda tangannya valid. Putusan integritas diukur dengan *false alarm*, yaitu proporsi citra utuh yang dinilai *malicious*, dan laju deteksi, yaitu proporsi citra termanipulasi yang dinilai *malicious*, dengan interval kepercayaan 95% Wilson. Pembanding adalah kondisi tanpa sistem, yaitu pencocokan *hash* SHA-256 berkas. Kualitas visual diukur dengan *PSNR* dan *SSIM*, akurasi lokalisasi tanpa kanal dengan *precision*, *recall*, *F1*, dan *IoU* pada tingkat blok terhadap *ground-truth mask*, dan pengklasifikasi dengan validasi silang 5-fold yang memisahkan sampel per citra sumber.

I. Model ancaman

Tabel 5 memetakan ancaman yang dipertimbangkan ke mekanisme penangkalnya dan ke bukti yang tersedia. Penyerang diasumsikan dapat mengubah piksel citra, menyalin blok di dalam citra yang sama, mengirim ulang citra lewat kanal yang mengkompresi ulang, dan memalsukan *sertifikat*, tetapi tidak memegang kunci privat penerbit.

Tabel 5. Model ancaman, mekanisme, dan bukti pada studi ini

Ancaman	Mekanisme	Bukti saat ini
Kompresi sah (JPEG, WhatsApp)	BER ≤ τ, status benign	Simulasi JPEG dan uji WhatsApp
Manipulasi piksel lokal	BER per blok > τ, lokalisasi	200 citra; 9/12 di WhatsApp Web
Copy-move atau *splicing*	*Position-salt*	Uji unit (pytest)
Pemalsuan atau penyangkalan penerbit	ML-DSA-44	Uji unit empat skenario adversarial
Metadata dan berkas pendamping hilang	enrollment_id robust di LL	WhatsApp nyata 70/70; JPEG simulasi
Citra tidak terkait	BER mendekati 50%, ditolak	Uji *true-negative*
Resize, foto ulang layar	Tidak ditangani	Batas desain

J. Peran ML-DSA, ancaman jangka panjang, dan batasan kepercayaan

Tanda tangan disimpan di server, bukan di piksel, sehingga ukuran tanda tangan ML-DSA-44 tidak menjadi kendala penyisipan dan pilihan algoritma tanda tangan tidak memengaruhi hasil eksperimen pada piksel (bagian Hasil dan Pembahasan). Karena itu kontribusi yang diuji secara empiris adalah lapisan *watermark* dan perilakunya pada kanal nyata, sedangkan ML-DSA-44 adalah pilihan desain untuk lapisan bukti. Alasan pemilihannya adalah keabsahan jangka panjang: citra medis dan dokumen identitas diarsipkan puluhan tahun, sedangkan tanda tangan RSA atau ECDSA yang dibuat hari ini dapat dipalsukan bila komputer kuantum yang relevan tersedia (*trust now, forge later*), dan NIST telah menyusun rencana transisi dari algoritma klasik ke standar pascakuantum [22]. Server terpercaya yang menyimpan sertifikat dan kunci publik tetap menjadi titik kepercayaan, sehingga tanda tangan pascakuantum tidak menghilangkan ketergantungan pada server, melainkan menjaga agar bukti yang dihasilkannya tidak usang. Biaya tanda tangan diukur pada satu laptop Windows (prosesor Intel, Python 3.11) dengan pustaka Python murni dilithium-py, tanpa optimasi. ML-DSA-44 menandatangani root hash 32 byte dalam 84,8 ms (median dari 200 ulangan; persentil ke-95 254 ms) dan memverifikasinya dalam 22,8 ms, dengan tanda tangan 2.420 byte dan kunci publik 1.312 byte. Sebagai pembanding, Ed25519 dan ECDSA P-256 pada pustaka *cryptography* menandatangani dalam 0,16 dan 0,10 ms dengan tanda tangan 64 dan 71 byte. *Enrollment* satu citra 1.024 piksel memerlukan 815 ms untuk penyisipan watermark (median dari 5 citra), sehingga penandatanganan ML-DSA-44 menambah sekitar 10% waktu *enrollment* dan verifikasinya terjadi di server, bukan di perangkat penerima. Selisih terhadap algoritma klasik karenanya berupa ukuran dan waktu server, bukan beban pada citra atau pada ketahanan terhadap kompresi. Perlu ditegaskan bahwa ML-DSA-44 mengesahkan catatan enrollment (root hash sertifikat), bukan piksel citra yang diterima: penghubung antara citra dan catatan itu adalah *enrollment_id* di piksel, sedangkan keutuhan piksel diputuskan oleh BER fingerprint yang tidak ditandatangani. Position-salt diturunkan dari koordinat blok, bukan dari kunci rahasia, sehingga penyerang yang mengetahui algoritma dapat menyisipkan ulang fingerprint pada citra yang dipalsukan. Mengikat fingerprint ke kunci rahasia penerbit adalah pekerjaan lanjutan. Berbeda dengan [25], yang menyisipkan tanda tangan ke dalam citra dengan DWT–SVD, kami tidak menyisipkan tanda tangan: tanda tangan ML-DSA-44 sebesar 2.420 byte (19.360 bit) jauh melampaui 48 bit yang dibawa lapisan robust (pengenal 32 bit dan CRC-16), dan kami tidak menguji apakah penyisipan sebesar itu tahan terhadap kompresi aplikasi pesan.

Beberapa aspek infrastruktur kepercayaan belum dicakup implementasi referensi dan dinyatakan sebagai batasan: (1) distribusi dan pengikatan kunci publik ke identitas penerbit, yang pada praktiknya memerlukan infrastruktur kunci publik (PKI) dengan sertifikat dari otoritas tepercaya, sedangkan implementasi saat ini memakai satu kunci institusi; (2) cap waktu tepercaya, misalnya melalui *Time-Stamp Protocol* [23], agar waktu penerbitan dapat dibuktikan, bersama pemeliharaan bukti jangka panjang seperti *Evidence Record Syntax* [24]; (3) pencabutan kunci atau sertifikat yang bocor; dan (4) ketersediaan server, karena penelusuran *enrollment_id* bergantung pada server yang aktif. Bila server tidak tersedia, lapisan piksel tetap dapat memberi putusan integritas secara *blind*, tetapi bukti asal tidak dapat diambil. Aspek-aspek ini tidak diklaim teratasi oleh artikel ini dan menjadi agenda desain berikutnya.

K. Strategi pencarian literatur

Klaim kebaruan kami dibatasi pada pencarian berikut: kami belum menemukan evaluasi skema watermark citra yang terhubung ke tanda tangan pascakuantum pada citra yang telah melewati kompresi aplikasi pesan instan. Pada 26 Oktober 2026 kami menjalankan tiga *query* pada Crossref REST API dan OpenAlex untuk publikasi sejak 2020, dengan 20 hasil teratas menurut relevansi per *query* pada setiap basis data: (i) *post-quantum signature image watermarking authentication*; (ii) *ML-DSA Dilithium digital image watermark*; dan (iii) *lattice-based signature medical image authentication watermark*. Pada OpenAlex kami menambahkan dua kueri: (iv) post-quantum digital signature image watermarking dan (v) quantum-resistant image authentication watermark Dilithium. Scopus, IEEE Xplore, dan Web of Science tidak dijalankan karena memerlukan akses institusi. Pencarian web umum dengan kombinasi kata kunci yang sama melengkapinya. Penyaringan dilakukan pada judul dan abstrak. Dari 60 posisi hasil Crossref tidak ada yang menanamkan atau mengikat tanda tangan pascakuantum pada skema *watermark* citra; yang muncul adalah implementasi ML-DSA pada protokol, perangkat keras, dan IoT, serta skema *watermark* citra yang memakai hash atau tanda tangan klasik. OpenAlex menampilkan tiga karya yang mendekati: [25] menyisipkan metadata dan tanda tangan klasik atau pascakuantum ke dalam citra dengan watermark hibrida DWT–SVD, menyimpan tanda tangan itu di blockchain, dan mengevaluasi efisiensi, konsumsi sumber daya, dan biaya gas; [26] membandingkan kinerja tanda tangan pascakuantum untuk menandatangani media; dan [27] menerapkan SPHINCS+, Dilithium, dan Falcon pada autentikasi citra medis. Pada [25], ketahanan watermark dinyatakan sebagai sifat teknik DWT–SVD tetapi tidak diukur: teks lengkapnya tidak memuat uji kompresi JPEG atau aplikasi pesan, PSNR, SSIM, maupun lokalisasi manipulasi per blok.

Abstrak [26] dan [27] melaporkan kinerja kriptografi atau penerapan tanda tangan dan tidak menyebut hal-hal tersebut; teks lengkap keduanya tidak kami telaah. Selain itu, satu bab tinjauan tentang watermarking tahan kuantum [20] membahas pertukaran antara imperceptibility, kapasitas sisip, dan biaya komputasi. Secara teoretis, [21] menunjukkan bahwa skema *watermark* yang tahan manipulasi, tak dapat dipalsukan, dan dapat dideteksi publik mungkin ada, tetapi sebagian komponennya belum dapat dibangun dengan teknologi sekarang.

Pencarian ini tidak menggantikan tinjauan sistematis: basis data berlangganan (Scopus, IEEE Xplore, Web of Science) belum ditelusuri dan *query* hanya memakai bahasa Inggris. Klaim kami karena itu berbunyi "pada sumber yang ditinjau" dan harus dibaca sebagai batas pencarian, bukan bukti ketiadaan.

3. Hasil dan Pembahasan (Results and Discussion)

Seluruh hasil berasal dari studi berskala terbatas: 60 citra untuk satu generasi pengiriman melalui WhatsApp Web, 10 citra untuk aplikasi seluler, 6 citra untuk sapuan JPEG, dan satu dataset 200 citra untuk lokalisasi tanpa kanal. Angka-angka tersebut indikatif dan tidak boleh dibaca sebagai klaim generalisasi.

A. Penelusuran asal pada kanal WhatsApp

Tabel 6. Hasil uji pengiriman WhatsApp Web pada 60 citra (satu generasi)

Metrik	Hasil	Keterangan
enrollment_id terbaca dari piksel	60/60	48 utuh dan 12 manipulasi
Tanda tangan ML-DSA-44 valid	60/60	diambil lewat enrollment_id
Hash SHA-256 cocok dengan berkas asli	0/48	*false alarm* alur hash 100%
Citra utuh dinilai *malicious*	0/48	CI 95%: 0–7,4%
Status citra utuh	11 *intact*, 37 *benign*	-
BER rata-rata citra utuh	0,027 ± 0,017	maksimum 0,074
Blok di atas ambang (citra utuh)	3,3%	1.149 dari 35.289 blok
Manipulasi dinilai *malicious*	9/12 (75%)	CI 95%: 47–91%
Selisih blok di atas ambang terhadap pasangan utuh	+2 hingga +4	pada 12 dari 12 citra

Tabel 6 menunjukkan bahwa $enrollment_id$ terbaca dari piksel pada seluruh 60 berkas hasil unduhan WhatsApp Web, dan tanda tangan ML-DSA-44 valid pada seluruhnya. Ini hasil utama penelitian: dari satu berkas gambar tanpa metadata, penerima dapat menelusuri sertifikat dan tanda tangan penerbit. Pada kondisi tanpa sistem, tidak satu pun dari 48 berkas utuh hasil WhatsApp berhash sama dengan aslinya, sehingga alur *hash* menandai seluruhnya sebagai palsu. Sistem yang diusulkan tidak menandai satu pun sebagai *malicious* (Tabel 7). Keunggulan tersebut dapat diperkirakan dari *avalanche effect*, tetapi baru kali ini diukur pada kanal nyata, dan baru pada satu generasi pengiriman.

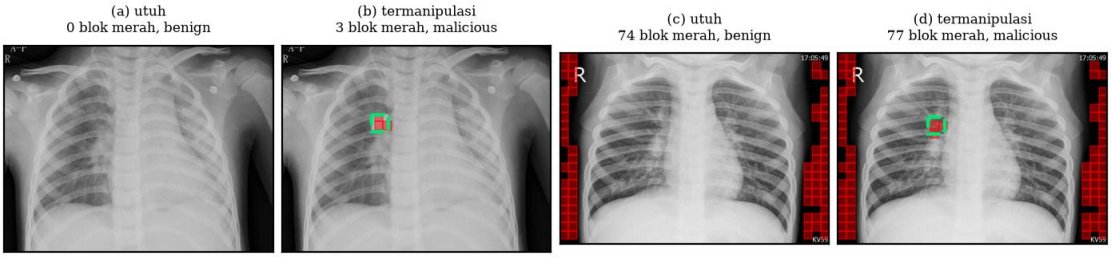
Tabel 7. Perbandingan alur tanpa sistem (hash SHA-256) dan dengan sistem pada berkas yang sama

Kondisi	Tanpa sistem	Dengan sistem	Status bukti
Citra sah setelah WhatsApp	Hash berbeda pada 48/48, ditandai palsu	Tidak ada yang *malicious* (0/48, WhatsApp Web)	Terukur, 48 citra
Manipulasi lokal	Tidak dibedakan dari kompresi, tanpa lokalisasi	Malicious 9/12, lokalisasi per blok	12 citra WhatsApp, 200 citra simulasi
Pembuktian penerbit	Tidak dapat dibuktikan	ML-DSA-44 lewat *enrollment_id*, 70/70	Terukur, 70 citra
Kebutuhan citra asli	Diperlukan	Tidak diperlukan (*blind*)	Desain dan uji

B. Deteksi dan lokalisasi manipulasi

Kompresi WhatsApp tetap meninggalkan jejak pada citra utuh: *BER* rata-rata 0,027, dan 34 dari 48 citra utuh memiliki setidaknya satu blok di atas ambang (hingga sekitar 11% blok pada citra tertentu). Jejak ini tidak dikenali sebagai manipulasi, tetapi *heatmap* menampilkan blok-blok tersebut, sehingga antarmuka sebaiknya menamainya blok di atas ambang, bukan blok termanipulasi. Temuan ini berbeda dari anggapan bahwa citra lolos WhatsApp tanpa distorsi seperti yang dilaporkan untuk skema spasial [15]: pada sistem ini distorsinya terukur pada setiap citra.

Deteksi manipulasi lokal adalah titik terlemah. Tempelan 64×64 piksel hanya memengaruhi sekitar empat blok, dan sembilan dari dua belas citra dinilai *malicious*; tiga lolos sebagai *benign*. Bila dibandingkan dengan pasangan utuhnya, jumlah blok di atas ambang naik 2 sampai 4 pada seluruh dua belas citra, sehingga sinyal manipulasi selalu ada, tetapi tenggelam dalam derau kompresi bila tidak ada pembanding. Dengan interval kepercayaan 47–91% pada n = 12, laju deteksi 75% belum dapat dianggap stabil.



Gambar 6. *Heatmap* blok di atas ambang (merah) setelah WhatsApp Web; kotak hijau menandai lokasi tempelan

Gambar 6 memperlihatkan dua pasangan citra. Pada pasangan (a, b), citra utuh tidak memiliki blok merah, sedangkan citra termanipulasi memiliki tiga blok merah tepat di dalam kotak tempelan sehingga lokasi manipulasi terlokalisasi dengan benar. Pada pasangan (c, d), citra utuh sudah memiliki 74 blok merah yang terkonsentrasi di bingkai hitam datar di tepi citra, sementara area anatomi hampir bebas. Tempelan pada (d) hanya menambah beberapa blok di dalam kotak hijau dan tenggelam di antara blok merah derau. Pola ini menunjukkan bahwa blok datar adalah sumber utama positif palsu tingkat blok setelah kompresi, sehingga mengabaikan atau membobot blok datar berpotensi memperbaiki lokalisasi; hal ini belum diuji dan menjadi pekerjaan lanjutan.

Satu catatan tentang citra utuh: berkas hasil *enrollment* yang tidak diubah sama sekali tidak selalu berstatus *intact*. Untuk mencari sumber *BER* dasar ini kami menjalankan ablasi pada 48 citra uji WhatsApp tanpa kompresi dan tanpa manipulasi, dengan empat kondisi: tanpa atau dengan lapisan ID, dan tanpa atau dengan pembulatan ke bilangan bulat 8-bit saat rekonstruksi. Tanpa pembulatan, *BER* dasar hampir nol (0,0010 tanpa ID; 0,00007 dengan ID), sehingga lapisan ID tidak membalik bit *fingerprint* pada *enrollment*, sesuai urutan penyisipannya. Dengan pembulatan 8-bit, *BER* rata-rata per citra menjadi 0,0057 tanpa ID, yang sesuai dengan kisaran 0,005–0,008 pada uji lokal kami dan identik antara kode yang memotong citra ke kelipatan 32 dan yang mempertahankan dimensi asli, lalu 0,0162 dengan ID (rentang 0,0003–0,0533 antarcitra; 0,128% blok melewati ambang). Selisih terkonsentrasi pada blok yang memuat piksel mendekati 0 atau 255: pada 12 citra yang diperiksa lebih lanjut, seluruh 807 blok ber-*BER* di atas 0,05 memuat piksel ≤ 8 atau ≥ 247 (rata-rata 0,115, dibandingkan 0,0004 pada blok lain). Penyebabnya adalah pemotongan (*clipping*) ke rentang 0–255 saat rekonstruksi, yang diperparah oleh pergeseran LL dari lapisan ID di latar gelap citra rontgen. Dengan demikian *BER* dasar bukan sifat *fingerprint*, melainkan akibat pembulatan dan *clipping*, dan pada berkas yang benar-benar dikirim (dengan ID) ia lebih tinggi daripada 0,005–0,008. Karena batas kelas *intact* berada dekat *BER* dasar itu, pembeda *intact* dan *benign* lemah, dan keduanya kami perlakukan sebagai tidak ada manipulasi.

C. Ketergantungan pada kanal dan kompresi

Tabel 8. Sapuan kompresi JPEG simulasi, resize, dan citra tak terkait pada layanan live (6 citra)

Perlakuan	Status	BER rata-rata	ID terbaca
Utuh (PNG)	6 *intact*	0,008	6/6
JPEG q90	6 *benign*	0,029	6/6
JPEG q80	6 *benign*	0,184	6/6
JPEG q70	6 *benign*	0,298	6/6
JPEG q60	6 *benign*	0,369	6/6
JPEG q50	3 *benign*, 3 *malicious*	0,412	6/6
JPEG q40	4 *malicious*, 2 *unrelated*	0,466	5/6
Tempel blok 64 px	6 *malicious*	0,011	6/6
Resize +0,3%	6 *unrelated*	0,467	0/6
Citra tak terkait	*unrelated*	0,500	tidak ada

Tabel 8 memberi skala pada kompresi WhatsApp. *BER* citra utuh setelah WhatsApp Web (0,027) setara dengan JPEG kualitas 90 (0,029), jadi kompresi WhatsApp Web pada uji ini tergolong ringan. Dengan pengklasifikasi akhir, zona *benign* terbentang sampai kualitas sekitar 60: pada kualitas 50 tiga dari enam citra dinilai *malicious*, dan pada kualitas 40 empat dari enam (dua citra tidak terbaca ID-nya sehingga dinilai *unrelated*). Batas yang lebih longgar ini menurunkan kepekaan terhadap manipulasi kecil pada kompresi berat (Tabel 9). Sebaliknya, *enrollment_id* bertahan lebih lama daripada putusan integritas: terbaca pada 6/6 citra sampai kualitas 50 dan 5/6 pada kualitas 40. Ini mendukung desain dua lapis, karena penelusuran ke penerbit masih mungkin ketika putusan piksel sudah tidak andal. BER citra utuh pada Tabel 8 (0,008;

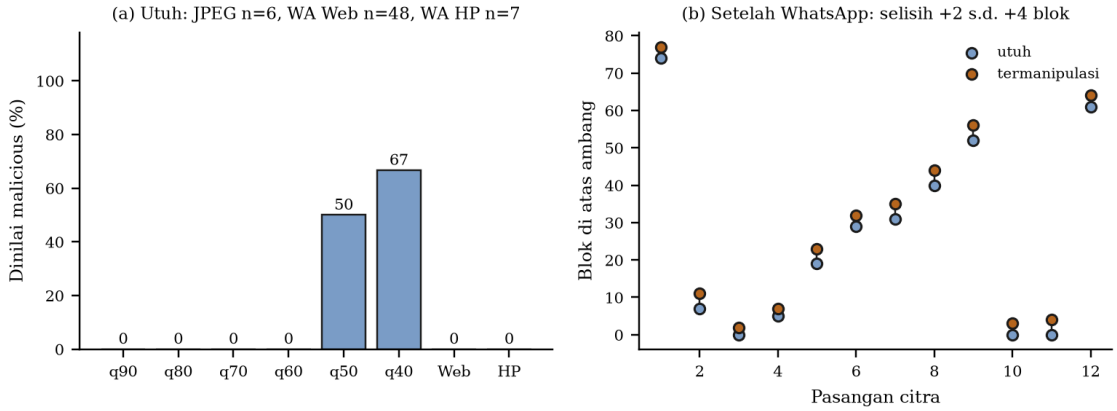
enam citra) lebih rendah daripada rerata ablasi pada 48 citra (0,016) karena komposisi citra, bukan perbedaan sistem: pada enam citra yang sama, skrip lokal menghasilkan rerata 0,0093 dan layanan live 0,0079, dengan selisih per citra di bawah 0,005.

Tabel 9. Perbandingan WhatsApp Web dan aplikasi seluler pada citra yang sama (7 utuh, 3 termanipulasi)

Metrik	WhatsApp Web	Aplikasi seluler
Ukuran berkas rata-rata (citra utuh)	222 KB	89 KB
BER rata-rata (citra utuh)	0,025	0,278 ± 0,012
Blok di atas ambang (citra utuh)	2,4%	18,7%
Citra utuh dinilai *malicious*	0/7	0/7 (CI 95%: 0–35%)
enrollment_id terbaca	10/10	10/10
Tanda tangan ML-DSA-44 valid	10/10	10/10
Citra termanipulasi dinilai *malicious*	1/3	0/3

Tabel 9 membandingkan kedua klien pada citra yang sama. Kompresi aplikasi seluler jauh lebih berat: berkas rata-rata 89 KB (WhatsApp Web 222 KB), *BER* citra utuh 0,278 (0,025), dan 18,7% blok berada di atas ambang (2,4%). *BER* ini mendekati *BER* JPEG kualitas 70 pada Tabel 8 (0,298), yaitu masih di dalam zona *benign* pengklasifikasi akhir, sehingga tidak satu pun dari 7 citra utuh dinilai *malicious* (CI 95%: 0–35%). Pengklasifikasi sebelumnya, yang dilatih tanpa lapisan ID, menilai 4 dari 7 citra utuh sebagai *malicious*; pada n = 7 selisih ini menunjukkan bahwa putusan pada BER setinggi ini sangat bergantung pada batas kelas. Hasil WhatsApp Web tidak boleh digeneralisasi ke aplikasi seluler.

Yang bertahan adalah lapisan *robust*: *enrollment_id* terbaca dan tanda tangan valid pada 10 dari 10 berkas seluler, walau putusan integritas tidak lagi andal. Pada kanal ini sistem masih dapat membuktikan asal-usul penerbit, tetapi belum dapat membedakan kompresi dari manipulasi kecil: tidak satu pun dari tiga citra termanipulasi dinilai malicious, sama seperti citra utuh, sehingga tidak adanya *false alarm* dicapai dengan mengorbankan deteksi (pengklasifikasi sebelumnya menilai 2 dari 3 citra termanipulasi *malicious*, sejalan dengan 4 dari 7 citra utuh). Temuan ini mengarah pada kebutuhan ambang yang dikalibrasi per kanal (misalnya τ adaptif-kanal atau pembeda berbasis sebaran blok), yang menjadi agenda utama pekerjaan lanjutan.



Gambar 7. Hasil uji: (a) citra utuh dinilai *malicious* menurut kualitas JPEG dan kanal; (b) blok di atas ambang pada pasangan utuh dan termanipulasi

Batas desain yang jelas adalah *resize*. *Resize* hanya 0,3% menggeser seluruh *grid* blok sehingga seluruh enam citra dinilai *unrelated* dengan *BER* 0,467 dan ID tidak terbaca, setara citra yang tidak terkait. Ini bukan galat, melainkan konsekuensi skema berbasis koordinat piksel eksak, dan kanal yang mengubah resolusi tidak tertangani oleh versi skema saat ini. Sinkronisasi *grid* setelah *resize* ringan menjadi pekerjaan lanjutan.

D. Lapisan robust pada kompresi simulasi

Tabel 10. Ekstraksi enrollment_id pada kompresi JPEG simulasi (9 citra 512×512)

Kondisi	ID benar terbaca	*False positive*	Beda PSNR
JPEG kualitas ≥ 60	100%	0	≈0,6 dB
JPEG kualitas 50	88,9%	0	≈0,6 dB

Tabel 10 menunjukkan bahwa identitas terbaca pada seluruh 9 citra untuk kualitas JPEG 60 ke atas dan pada 88,9% citra pada kualitas 50, tanpa *false positive*, dengan penurunan *PSNR* rata-rata sekitar 0,6 dB akibat lapisan tambahan. Kelayakan ini terkonfirmasi pada kanal nyata, yaitu 70 dari 70 berkas WhatsApp (Tabel 6 dan 9). Pengiriman berulang (*forward*) belum diuji.

E. Lokalisasi dan kualitas visual tanpa kanal

Tabel 11. Hasil kualitas visual, lokalisasi, dan klasifikasi pada 200 citra Chest X-Ray

Metrik	Hasil	Keterangan
PSNR (dB)	32,94 ± 0,16	rentang 32,46–33,43 (n = 200)
SSIM	0,723 ± 0,022	rentang 0,662–0,832
Precision	1,0000	TP 194.356, FP 0
Recall	0,99995	FN 9 (1.996 kasus manipulasi)
F1 / IoU	0,99998 / 0,99995	tingkat blok
Akurasi *Random Forest*	94,87%	5-fold per citra sumber, n = 5.200

Tanpa kanal, lokalisasi sangat akurat (Tabel 11): dari 1.996 kasus manipulasi, tidak ada blok utuh yang ditandai manipulasi (FP = 0) dan hanya 9 blok manipulasi yang terlewat dari lebih dari 194 ribu. Hasil ini wajar untuk kepadatan penyisipan tinggi karena setiap blok membawa *fingerprint* 256 bit sendiri, sehingga tidak ada blok buta. Hasil ini menunjukkan bahwa kelemahan deteksi pada WhatsApp berasal dari kompresi kanal, bukan dari skema itu sendiri.

Harga yang dibayar adalah kualitas visual. *PSNR* 32,9 dB dan *SSIM* 0,72 jauh di bawah skema pembanding Hacini dkk. [11] yang mencapai 44,2 dB dan *SSIM* di atas 0,98. Penyebab yang paling mungkin adalah perbedaan kepadatan penyisipan: skema pembanding menyebar 256 bit total ke seluruh blok 8×8 pada citra 512×512 (rata-rata sekitar 0,06 bit per blok), sedangkan skema ini menyisipkan 256 bit penuh pada setiap blok, ditambah Q_EMBED = 16 yang belum dikalibrasi seketat itu. Kesenjangan ini kami anggap sebagai keterbatasan utama dan kandidat eksperimen sensitivitas berikutnya.

Random Forest mencapai akurasi 94,87% pada validasi silang 5-fold yang memisahkan sampel per citra sumber (*GroupKFold*), dengan 5.200 sampel dari 200 citra sumber (1.200 *intact*, 1.800 *benign*, 2.200 *malicious*) yang seluruhnya memuat lapisan *enrollment_id*. *Recall intact* 96,6% (1.159 dari 1.200); hanya 5 dari 2.200 kasus malicious yang lolos, dan 3 dari 3.000 kasus tanpa manipulasi yang salah dinilai malicious. Kesalahan terbanyak adalah *benign* yang terbaca *intact* (218 dari 1.800), sehingga pembeda *intact* dan *benign* lemah dan keduanya diperlakukan sebagai tidak ada manipulasi. Angka ini lebih rendah daripada 98,97% pada versi pertama (3.996 sampel tanpa lapisan ID dan tanpa pemisahan per citra sumber), yang terlalu optimistis karena varian dari citra yang sama muncul di data latih dan uji.

F. Implikasi praktis

Bagi institusi medis, pemisahan peran penerbit (memegang kunci privat) dan verifikator (hanya kunci publik) memungkinkan audit citra tanpa memperluas permukaan serangan pada kunci. Bagi penyedia e-KYC, verifikasi *blind* berarti tidak ada arsip citra asli yang perlu dijaga, yang mengurangi beban penyimpanan dan risiko kebocoran data pribadi. Lapisan *enrollment_id* menjawab skenario yang paling lazim di lapangan, yaitu citra yang beredar sebagai satu berkas tunggal lewat aplikasi perpesanan, dan pengenal itu terbukti terbaca pada seluruh 70 citra uji WhatsApp, meskipun pengiriman berulang belum diuji.

G. Ancaman terhadap validitas

Validitas internal terancam oleh kalibrasi parameter yang dilakukan pada data yang sama dengan evaluasi dan oleh sampel yang diperbanyak dari citra sumber yang sama. Validitas eksternal terancam oleh satu dataset rontgen dada, satu generasi pengiriman, dan hanya 10 citra pada aplikasi seluler, sedangkan perilaku kompresi aplikasi dapat berubah antarversi. Validitas konstruk terancam oleh *PSNR* dan *SSIM* yang hanya mengukur kemiripan terhadap citra asli, bukan dampak klinis. Untuk menekan ancaman ini, penelitian lanjutan akan memakai beberapa modalitas, pemisahan data latih dan uji per citra sumber, serta pengujian WhatsApp pada beberapa perangkat dan versi aplikasi.

H. Keterbatasan

Keterbatasan penelitian ini adalah: (1) validasi masih pada satu dataset dan 200 citra sumber, jauh di bawah skala pembanding; (2) ambang $\tau = 0{,}30$ masih hasil kalibrasi manual dan belum diturunkan dengan *ROC* dan indeks Youden; (3) belum ada perbandingan kuantitatif langsung terhadap skema lain pada data yang sama; (4) uji WhatsApp baru satu generasi pengiriman, dengan 12 sampel termanipulasi

pada WhatsApp Web dan hanya 10 citra pada aplikasi seluler; (5) laju deteksi manipulasi kecil (empat blok) baru 75% dengan interval kepercayaan lebar; (6) *BER* dasar citra utuh sekitar 0,6–1,6% (akibat pembulatan 8-bit dan *clipping* piksel, lihat ablasi pada Subbab B) melemahkan pembeda *intact* dan *benign*; (7) pada aplikasi seluler tidak ada *false alarm* (0/7) tetapi juga tidak ada deteksi (0/3), karena putusan piksel pada BER sekitar 0,28 tidak lagi memisahkan kompresi dari manipulasi kecil; dan (8) *resize* serta foto ulang layar tidak tertangani.

I. Agenda eksperimen lanjutan

Agar klaim dapat diperkuat, pengujian skala penuh direncanakan pada tiga modalitas, yaitu rontgen dada, OCT retina, dan MRI otak, dengan skala yang sebanding dengan pembanding utama. Ambang *BER* akan diturunkan dari kurva *ROC* dan indeks Youden untuk tiap kanal (WhatsApp Web dan seluler), dan pengaruh pembobotan blok datar terhadap *false alarm* akan diuji. Pengklasifikasi akan dilatih dan diuji dengan sampel yang meniru kompresi aplikasi seluler, agar kepekaan terhadap manipulasi kecil pada kompresi berat dapat dinilai. Perbandingan kuantitatif langsung terhadap skema Hacini dkk. [11] pada data yang sama akan mencakup *PSNR*, *SSIM*, *BER*, dan tabel ketahanan. Eksperimen sensitivitas kepadatan penyisipan terhadap Q_EMBED akan mengukur *trade-off* antara kualitas visual dan akurasi lokalisasi. Uji WhatsApp diperluas ke beberapa perangkat dan versi aplikasi dengan pengiriman berulang, termasuk keterbacaan *enrollment_id* dan laju *false alarm* pada tiap generasi. Strategi penyisipan yang sadar-*clipping*, misalnya melewati atau menggeser blok yang memuat piksel jenuh, juga akan diuji karena ablasi menunjukkan *clipping* sebagai sumber BER dasar.

4. Kesimpulan (Conclusion)

Artikel ini menunjukkan bahwa citra yang dikirim melalui WhatsApp dapat diperiksa asal dan keutuhannya dari satu berkas tunggal dengan skema *watermark* ganda pada domain *DWT* yang terhubung ke tanda tangan pascakuantum ML-DSA-44. Pada pengiriman WhatsApp Web terhadap 60 citra, $enrollment_id$ terbaca dan tanda tangan valid pada seluruhnya, tidak ada citra utuh yang dinilai *malicious* (0/48) sementara alur *hash* menolak seluruhnya, dan manipulasi lokal terdeteksi pada 9 dari 12 citra. Pada aplikasi seluler yang kompresinya lebih berat (*BER* sekitar 0,28), $enrollment_id$ tetap terbaca pada 10 dari 10 citra, tidak ada citra utuh yang dinilai malicious (0/7) tetapi tiga citra termanipulasi juga tidak terdeteksi (0/3), sehingga pada kanal tersebut sistem membuktikan asal tetapi belum memeriksa keutuhan. Dengan demikian penelusuran asal andal di kedua kanal, sedangkan deteksi manipulasi baru terukur pada kompresi ringan (WhatsApp Web, 9 dari 12 citra) dan memerlukan kalibrasi per kanal di atasnya. Derau tingkat blok menumpuk di area datar, dan *resize* tetap menjadi batas desain. Pada 200 citra *Chest X-Ray* tanpa kanal, skema mencapai presisi lokalisasi 1,0000 dan *recall* 0,99995 dengan *PSNR* 32,9 dB yang lebih rendah dari skema pembanding. Pekerjaan lanjutan mencakup perluasan dataset, kalibrasi ambang per kanal dengan *ROC* dan Youden, pembobotan blok datar, pelatihan pengklasifikasi dengan sampel kompresi seluler, pengiriman WhatsApp berulang pada aplikasi web dan seluler, perbaikan kualitas visual melalui kalibrasi kepadatan penyisipan, penajaman pembeda *intact* dan *benign*, serta sinkronisasi *grid* untuk *resize* ringan. Pilihan ML-DSA-44 tidak memengaruhi hasil eksperimen pada piksel; nilainya terletak pada keabsahan bukti jangka panjang, dengan infrastruktur kunci publik, cap waktu, pencabutan, dan ketersediaan server sebagai batasan yang belum dicakup.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada dosen pembimbing dan Program Studi Teknik Informatika S-2 Universitas Pamulang atas arahan selama penelitian ini.

Pernyataan Etik dan Deklarasi Artikel

Penggunaan Kecerdasan Artifisial: Penulis menggunakan asisten AI (Claude, Anthropic) untuk membantu menyusun draf naskah dan kode pendukung. Seluruh isi, data, dan referensi telah ditinjau dan diverifikasi oleh penulis.

Kontribusi Penulis (Author Contribution Statement): Alif Maulana Arifin: penulisan naskah awal, pengumpulan data, analisis, dan visualisasi. Dedy Suhariyanto: konseptualisasi dan metodologi. Ferdiansyah Winardi: peninjauan (review) dan penyuntingan naskah. Seluruh penulis menyetujui naskah akhir.

Pendanaan (*Funding Statement*): Tidak ada pendanaan dari pihak manapun, hanya internal saja

Konflik Kepentingan (Conflict of Interest Statement): Penulis menyatakan tidak terdapat konflik kepentingan dalam penelitian dan penerbitan artikel ini.

Daftar Pustaka

[1] European Society of Radiology (ESR), "Usability of irreversible image compression in radiological imaging. A position paper by the European Society of Radiology (ESR)," *Insights Imaging*, vol. 2, no. 2, pp. 103-115, 2011, doi: 10.1007/s13244-011-0071-x.

[2] D. Koff *et al.*, "Pan-Canadian evaluation of irreversible compression ratios ("lossy" compression) for development of national guidelines," *J. Digit. Imaging*, vol. 22, no. 6, pp. 569-578, 2009, doi: 10.1007/s10278-008-9139-7.

[3] N. Soni, "Forensic value of Exif data: An analytical evaluation of metadata integrity across image transfer methods," *Perspect. Legal Forensic Sci.*, vol. 2, no. 2, p. 10006, 2025, doi: 10.70322/plfs.2025.10006.

[4] D. Babbar, A. Jain, A. Sharma, and D. Banerjee, "Forgery detection in Aadhaar-based KYC using deep learning," in *Proc. 16th Int. Conf. Mach. Learn. Comput.*, 2024, doi: 10.1145/3651671.3651691.

[5] Y. Mirsky, T. Mahler, I. Shelef, and Y. Elovici, "CT-GAN: Malicious tampering of 3D medical imagery using deep learning," in *Proc. 28th USENIX Security Symp.*, 2019, pp. 461-478.

[6] P. W. Wong, "A public key watermark for image verification and authentication," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 1, 1998, pp. 455-459, doi: 10.1109/ICIP.1998.723526.

[7] P. W. Wong and N. Memon, "Secret and public key image watermarking schemes for image authentication and ownership verification," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1593-1601, 2001, doi: 10.1109/83.951543.

[8] D. Shullani, M. Fontani, M. Iuliani, O. Al Shaya, and A. Piva, "VISION: A video and image dataset for source identification," *EURASIP J. Inf. Secur.*, vol. 2017, art. 15, 2017, doi: 10.1186/s13635-017-0067-2.

[9] J. J. Eggers and B. Girod, "Blind watermarking applied to image authentication," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2001, doi: 10.1109/icassp.2001.941335.

[10] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Academic Press, 1999.

[11] I. Hacini *et al.*, "Quantum-inspired post-quantum secure DCT watermarking for tamper detection and localization in medical images," *Array*, vol. 30, p. 100870, 2026, doi: 10.1016/j.array.2026.100870.

[12] R. Taj *et al.*, "Lightweight dual-watermarking framework for medical image authentication and integrity preservation," *Sci. Rep.*, vol. 16, 2025, doi: 10.1038/s41598-025-30615-z.

[13] M. Cedillo-Hernandez, A. Cedillo-Hernandez, F. J. Garcia-Ugalde, and J. C. Sanchez-Garcia, "Semi-fragile watermarking scheme for high-resolution color images: Tamper identification, ownership authentication, and self-recovery," *Algorithms*, vol. 19, no. 1, p. 28, 2025, doi: 10.3390/a19010028.

[14] X. Li, Q. Chen, R. Chu, and W. Wang, "Block mapping and dual-matrix-based watermarking for image authentication with self-recovery capability," *PLOS ONE*, 2024, doi: 10.1371/journal.pone.0297632.

[15] S. H. Basha and B. Jaison, "A novel secured Euclidean space points algorithm for blind spatial image watermarking," *EURASIP J. Image Video Process.*, vol. 2022, 2022, doi: 10.1186/s13640-022-00590-w.

[16] A. A. Yavuz, S. Darzi, and S. E. Nouma, "LiteQSign: Lightweight and quantum-safe signatures for heterogeneous IoT applications," arXiv:2311.18674, 2023.

[17] National Institute of Standards and Technology, "Module-Lattice-Based Digital Signature Standard," FIPS 204, U.S. Department of Commerce, 2024, doi: 10.6028/NIST.FIPS.204.

[18] A. Diwan and A. K. Roy, "Detection and localization of copy-move tampering along with adversarial attack in a digital image," *Discover Comput.*, vol. 28, no. 1, p. 136, 2025, doi: 10.1007/s10791-025-09658-3.

[19] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, 2018, doi: 10.1016/j.cell.2018.02.010.

[20] N. Varshney, J. Saxena, A. Mittal, and S. Gupta, "Quantum-resistant digital watermarking: Preparing for post-quantum cryptography," in *Digital Watermarking in Cloud Environments for Copyright Protection*. IGI Global Scientific Publishing, 2025, pp. 229-244, doi: 10.4018/979-8-3373-3785-2.ch010.

[21] J. Fairoze, G. Ortiz-Jimenez, M. Vecerik, S. Jha, and S. Gowal, "On the difficulty of constructing a robust and publicly-detectable watermark," *arXiv:2502.04901*, 2025.

[22] D. Moody, R. Perlner, A. Regenscheid, A. Robinson, and D. Cooper, "Transition to post-quantum cryptography standards," National Institute of Standards and Technology, NIST IR 8547 (Initial Public Draft), Nov. 2024, doi: 10.6028/NIST.IR.8547.ipd.
[23] C. Adams, P. Cain, D. Pinkas, and R. Zuccherato, "Internet X.509 public key infrastructure time-stamp protocol (TSP)," RFC 3161, IETF, Aug. 2001.
[24] T. Gondrom, R. Brandner, and U. Pordesch, "Evidence record syntax (ERS)," RFC 4998, IETF, Aug. 2007.
[25] M. Alkhatib, "A multifaceted deepfake prevention framework integrating blockchain, post-quantum cryptography, hybrid watermarking, human oversight, and policy governance," *Computers*, vol. 14, no. 11, art. 488, 2025, doi: 10.3390/computers14110488.
[26] M. Alkhatib, "Performance evaluation of NIST-standardized post-quantum and symmetric ciphers for mitigating deepfakes," *Cryptography*, vol. 10, no. 2, art. 15, 2026, doi: 10.3390/cryptography10020015.
[27] K. S. Roy et al., "Post-quantum digital signatures for enhanced medical image security," *IET Quantum Commun.*, vol. 6, no. 1, art. e70006, 2025, doi: 10.1049/qtc2.70006.