Journal of Computer Science and Information Technology (JCSIT)
Volume 3, No 3 – Juni 2026
e-ISSN : 3031-8467

DETEKSI TIMPA GAMBAR DENGAN TEKS MENGGUNAKAN CONVOLUTIONAL NEURAL NETWORK DAN OPTICAL CHARACTER RECOGNITION

Muhammad Alif Syafan[1], Fahrim Irhamna Rachman[2], Darniati[3]
[1,2,3]Universitas Muhammadiyah Makassar, Indonesia
[1]105841114422@student.unismuh.ac.id [2]fachrim141020@unismuh.ac.id
[3]darniati@unismuh.ac.id
* Coresponding Author

Received: 03-06- 2026	Revised: 20-06-2026	Approved: 27-06-2026

ABSTRAK

Penyebaran disinformasi melalui manipulasi timpa gambar dengan teks semakin marak dan sulit dideteksi karena artefak visualnya yang sangat halus dan rentan tersamarkan oleh kompresi citra. Penelitian ini mengusulkan arsitektur multi-stream network multimodal untuk mendeteksi manipulasi tersebut. Kebaruan (novelty) dan kontribusi utama penelitian ini terletak pada integrasi adaptif antara analisis anomali visual tingkat piksel dan anomali tipografi menggunakan gated attention mechanism. Metode usulan menggabungkan Convolutional Neural Network (CNN) berarsitektur EfficientNet-B3 yang dipertajam Sobel filter sebagai pengekstraksi fitur visual, dengan Optical Character Recognition (OCR) berbasis EasyOCR sebagai pengekstraksi fitur statistik teks. Evaluasi terhadap dataset Real Text Manipulation (RTM) membuktikan bahwa model multimodal usulan lebih unggul dibandingkan model CNN tunggal. Pada ambang batas optimal 0,4961, model mencapai akurasi 73,18% dan F1-score 0,7959. Integrasi OCR terbukti sangat efektif menambal kelemahan deteksi visual murni dengan lonjakan tingkat recall mencapai 79,05%. Meskipun pengujian generalisasi pada dataset media sosial dan generative AI masih mengalami penurunan akibat domain gap kompresi ekstrem, integrasi CNN dan OCR secara signifikan berkontribusi meningkatkan sensitivitas deteksi manipulasi teks untuk keperluan forensik digital.

Kata Kunci: *Timpa Teks, Manipulasi Citra, Convolutional Neural Network (CNN), Optical Character Recognition (OCR), Multi-stream Network*

PENDAHULUAN

Pemanfaatan platform media sosial sebagai media penyebaran informasi membawa dampak positif sekaligus negatif. Platform seperti Facebook memungkinkan pengguna membuat dan membagikan berbagai bentuk konten digital, termasuk teks dan gambar. Penelitian terdahulu pada tahun 2023 oleh Yang menunjukkan bahwa konten visual seperti gambar dan konten visual satir sering digunakan dalam penyebaran misinformasi di Facebook karena visual cenderung lebih menarik dan lebih mudah dipercaya oleh pengguna dibandingkan teks saja [1]. Selain itu, kajian linguistik forensik pada tahun 2025 oleh Putra Surya menyoroti bahwa interaksi pengguna Facebook juga menunjukkan adanya berbagai bentuk penyimpangan komunikasi di ruang digital seperti provokasi, fitnah, dan manipulasi pesan dalam unggahan pengguna [2]. Kondisi ini menunjukkan bahwa media sosial, khususnya Facebook, rentan terhadap penyebaran konten yang dimanipulasi. Kerentanan tersebut dibuktikan dengan tingginya peredaran informasi palsu di ruang digital Indonesia. Berdasarkan data dari Kementerian Komunikasi dan Digital (Komdigi), terdapat 1.615 isu hoaks yang ditangani di berbagai platform digital sepanjang tahun 2023 [3]. Angka tersebut mengalami peningkatan pada tahun 2024, di mana Komdigi mengidentifikasi sebanyak 1.923 konten hoaks [4]. Dalam ekosistem penyebaran disinformasi ini, temuan dari pemeriksa fakta

Masyarakat Antifitnah Indonesia (Mafindo) mengonfirmasi bahwa modus yang sangat sering dimanfaatkan di media sosial adalah manipulasi visual, khususnya modifikasi gambar dan tangkapan layar (*screenshot*) pesan berantai menggunakan teks palsu yang tidak sesuai dengan fakta aslinya [5].

Salah satu bentuk manipulasi visual yang sering ditemukan tersebut adalah manipulasi timpa gambar dengan teks, yaitu proses menimpa atau mengintervensi suatu citra asli menggunakan elemen teks baru sehingga menghasilkan informasi visual yang berbeda dari konteks aslinya. Sebuah penelitian pada tahun 2024 oleh Zhu menjelaskan bahwa manipulasi citra digital merupakan salah satu bentuk pemalsuan visual yang semakin banyak ditemukan pada konten internet dan media sosial [6]. Modus manipulasi ini sering muncul pada tangkapan layar percakapan, *headline* berita, maupun konten humor internet yang beredar luas di platform media sosial. Manipulasi timpa gambar dengan teks juga berkaitan dengan fenomena disinformasi multimodal di ruang digital yang menggabungkan teks dan gambar. Menyikapi fenomena ini, riset pada tahun 2025 oleh Jadhav mengidentifikasi di mana pendekatan deteksi berbasis teks saja sering tidak cukup untuk mengidentifikasi manipulasi informasi pada konten multimodal karena informasi dapat dimanipulasi melalui hubungan antara visual dan tekstual [7].

Dalam bidang forensik citra, penelitian manipulasi gambar umumnya berfokus pada teknik manipulasi umum seperti *image splicing* dan *copy-move* yang bertujuan mengubah isi citra digital. Berbeda dengan manipulasi objek pada citra, manipulasi timpa gambar dengan teks memiliki karakteristik yang lebih spesifik karena artefaknya sering muncul secara halus pada batas karakter atau *outline* huruf. Temuan dari sebuah penelitian pada tahun 2025 oleh Bae mengungkapkan fakta bahwa artefak manipulasi pada citra yang mengandung teks sering berada pada area tepi karakter sehingga sulit dideteksi secara visual dan dapat tersamarkan oleh proses kompresi citra [8]. Untuk menangkap anomali tersebut, sebuah pengujian sistem pada tahun 2025 oleh Wang membuktikan bahwa metode *convolutional neural network* (CNN) menjadi pendekatan yang tangguh dalam mengekstraksi fitur visual tingkat piksel dan mendeteksi inkonsistensi tekstur yang tidak terlihat pada analisis citra global [9]. Selain tantangan deteksi, keterbatasan dataset juga menjadi masalah dalam penelitian forensik citra ini. Sebagai solusi atas keterbatasan ini, sebuah penelitian pada tahun 2025 oleh Luo memperkenalkan dataset *real text manipulation* (RTM) yang merepresentasikan kondisi manipulasi yang lebih realistis, sekaligus menunjukkan bahwa banyak metode deteksi mengalami penurunan performa ketika diuji pada citra dunia nyata [10].

Berdasarkan penelitian terdahulu, deteksi manipulasi dokumen dan konten media sosial berkembang melalui pendekatan fitur visual, OCR, dan pembelajaran multimodal. Integrasi fitur visual dan tekstual terbukti mampu meningkatkan performa klasifikasi dibandingkan pendekatan unimodal [11]. Arsitektur *multi-stream* juga menunjukkan kemampuan yang baik dalam memproses berbagai fitur secara paralel sehingga mampu mendeteksi anomali secara lebih optimal.

Tabel 1. Penelitian Terkait dan *Research Positioning*

Referensi	Fokus	Pendekatan	Keterbatasan
Bae et al.	Manipulasi dokumen (*edge-focused*)	Unimodal (Visual CNN)	Rentan terhadap *blending* halus pada batas karakter.
Wang et al.	Pemalsuan dokumen identitas	Unimodal (CNN *Texture*)	Mengabaikan informasi tipografi dan struktur teks.
Luo et al.	Teks pada citra dunia nyata (RTM)	Unimodal (Visual dan *Frequency*)	Kesulitan membedakan artefak teks palsu dengan *noise* kompresi.
Jadhav et al.	Berita palsu	Multimodal (NLP dan Visual)	Berfokus pada makna semantik, bukan forensik piksel teks.
El-amrany et al.	Klasifikasi meme berbahaya	Multimodal (Visual dan Teks)	Tidak dirancang spesifik untuk deteksi artefak manipulasi teks.
Chen et al.	Manipulasi citra secara umum	MVSS-Net (Multi-View Multi-Scale)	Tidak dirancang untuk mengekstrak anomali tipografi teks.
Guillaro et al.	Pemalsuan citra dan lokalisasi	TruFor (RGB dan *Noiseprint*)	Mengabaikan fitur statistik susunan dan tata letak karakter.
Bui et al.	Deteksi *deepfake*	*Dual-branch Spatial-Frequency*	Berfokus pada wajah atau objek.
Penelitian ini	Timpa teks pada gambar	Multimodal (CNN dan OCR)	Menggabungkan ekstraksi anomali piksel visual dan statistik tipografi secara paralel.

Berdasarkan tinjauan pada Tabel 1, terlihat adanya kesenjangan penelitian (*research gap*). Berbagai metode deteksi forensik telah dikembangkan untuk mendeteksi dan melokalisasi manipulasi citra. Sebagai contoh, MVSS-Net mengombinasikan *multi-view* dan *multi-scale supervision* untuk mengeksploitasi karakteristik distribusi *noise* serta artefak pada batas wilayah manipulasi sehingga mampu meningkatkan deteksi dan lokalisasi manipulasi citra [12]. Selanjutnya, Guillaro dkk. memperkenalkan TruFor yang mengintegrasikan *low-level clues* berupa sidik jari (*fingerprint*) yang sensitif terhadap *noise* dengan *high-level visual clues* dari citra RGB untuk menghasilkan deteksi dan lokalisasi pemalsuan citra yang lebih andal [13]. Selain itu, *lightweight dual-branch spatial-frequency* menunjukkan bahwa pemanfaatan fitur spasial dan frekuensi secara bersamaan mampu meningkatkan kinerja deteksi sekaligus lokalisasi manipulasi pada kasus *deepfake* [14].

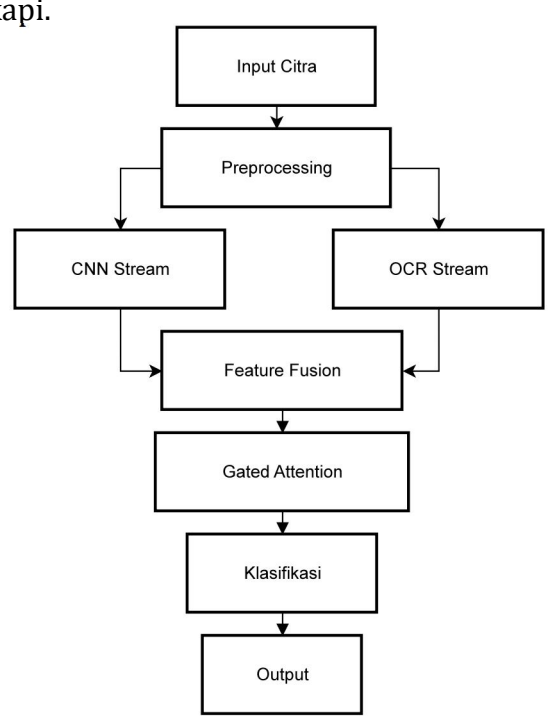
Berdasarkan kesenjangan tersebut, penelitian ini mengusulkan pendekatan multimodal berbasis *multi-stream network* untuk mendeteksi manipulasi timpa gambar dengan teks menggunakan metode *convolutional neural network* (CNN) untuk menganalisis konsistensi fitur visual citra, serta *optical character recognition* (OCR) untuk mengekstraksi informasi tekstualnya secara paralel. Secara eksplisit, kontribusi ilmiah (*scientific contribution*) dari penelitian ini adalah:

1. Mengintegrasikan model CNN (EfficientNet-B3) yang dipertajam Sobel filter dengan pustaka EasyOCR untuk menganalisis anomali visual dan statistik teks.
2. Menggunakan *gated attention mechanism* untuk menyeimbangkan dependensi bobot antara fitur visual dan tekstual secara otomatis.

3. Mengevaluasi ketahanan model secara langsung pada citra media sosial dan manipulasi dari *generative AI*.

METODE PENELITIAN

Dalam rancangan sistem ini, pemrosesan dirancang untuk memiliki peran yang saling melengkapi.



Gambar 1. Flowchart Penelitian

Proses dimulai dari tahap pra-pemrosesan di mana citra diseragamkan ukurannya (*resizing*), distabilkan intensitas warnanya melalui normalisasi ke format desimal, serta ditingkatkan variasinya melalui augmentasi data. Setelah itu, sistem membagi proses menjadi dua jalur utama yang berjalan paralel. Jalur pertama adalah *CNN stream*, di mana gambar diproses melalui *edge filter* (operator Sobel) untuk memperjelas artefak tepi hasil pengeditan. Pola fitur visual kemudian diekstraksi menggunakan arsitektur EfficientNet-B3 sebagai *backbone* untuk menghasilkan vektor fitur berdimensi 256. Penggunaan EfficientNet memungkinkan model menangkap fitur hierarkis yang lebih kompleks secara efisien. Secara paralel, jalur kedua yaitu *OCR stream* beroperasi sebagai sensor anomali struktur tipografi menggunakan pustaka EasyOCR. Proses ini mengekstrak lima dimensi fitur statistik mentah dari setiap blok teks, yaitu rata-rata probabilitas prediksi, variansi posisi vertikal, kepadatan deteksi, serta nilai probabilitas maksimal dan minimal. Vektor fitur OCR ini dilewatkan pada lapisan *batch normalization 1D* dan diekspansi oleh *dense layer* untuk menghasilkan vektor fitur tekstual berdimensi 128.

Setelah fitur diekstraksi secara paralel, penyatuan modalitas (*feature fusion*) dilakukan melalui teknik *concatenation*. Vektor visual dan tekstual disambungkan secara linear untuk menghasilkan vektor fitur gabungan multimodal berdimensi 384 seperti yang ditunjukkan pada persamaan berikut:

$$F_{combined} = [F_{visual_{sobel}} \oplus F_{ocr}]$$

Vektor gabungan ini tidak langsung masuk ke modul klasifikasi akhir, melainkan dilewatkan pada lapisan *gated attention layer*. Vektor diproses oleh lapisan *dense* dengan fungsi aktivasi sigmoid untuk menghasilkan *gate weights*:

$$g = \sigma(W_a \cdot F_{combined} + b_a)$$

Vektor bobot ini kemudian dikalikan secara langsung dengan fitur gabungan menggunakan operasi perkalian Hadamard (*element-wise multiplication*) guna menyeimbangkan dependensi fitur visual dan tekstual, menghasilkan fitur terboboti final:

$$F_{gated} = F_{combined} \odot g$$

Terakhir, fitur terboboti final diklasifikasikan menggunakan *fully connected layer* berbasis aktivasi sigmoid untuk menentukan apakah gambar tersebut asli atau hasil manipulasi:

$$y = f(F_{gated})$$

Untuk menjamin reproduktibilitas penelitian, seluruh inisialisasi bobot dan pembagian data dikunci menggunakan nilai *fixed random seed* sebesar 42. Seluruh eksperimen dijalankan menggunakan akselerasi *graphics processing unit* (GPU) dengan bahasa pemrograman Python 3 dan *framework* PyTorch. Citra *inputan* disamakan ukurannya menjadi 224x224 piksel dan dilengkapi dengan augmentasi rotasi dan *horizontal flip*. Pelatihan model dioptimalkan menggunakan algoritma AdamW dengan *batch size* 32. Pelatihan arsitektur multimodal usulan dirancang dalam dua tahap, yakni 2 *epoch* untuk *warmup* dengan *learning rate* 1e-3 (parameter *backbone* dibekukan), diikuti oleh 8 *epoch* untuk *fine-tuning* menggunakan *differential learning rate* (5e-5 pada fitur CNN dan 5e-4 pada ekstraktor OCR). *Loss function* mengandalkan *binary cross entropy with logits loss* (BCEWithLogitsLoss) yang dilengkapi dengan pembobotan kelas sebesar 0,5806 untuk meminimalkan bias klasifikasi pada dataset yang tidak seimbang.

Tahap evaluasi arsitektur dilakukan menggunakan pendekatan *ablation study* yang bertujuan menganalisis kontribusi masing-masing komponen terhadap performa akhir model. Pengujian ini dibagi ke dalam empat skenario, yakni skenario *baseline* (CNN murni), skenario visual dengan filter tepi (CNN dan Sobel), skenario OCR tunggal, dan skenario arsitektur lengkap usulan penelitian (CNN, Sobel, OCR, dan Attention). Evaluasi dilakukan melalui analisis kuantitatif dan kualitatif. Analisis kuantitatif mengukur performa model secara objektif menggunakan *confusion matrix* untuk menghitung akurasi, presisi, *recall*, dan *F1-score*. Sementara itu, analisis kualitatif berfokus pada inspeksi visual terhadap hasil prediksi model untuk mengidentifikasi penyebab kesalahan.

HASIL PENELITIAN DAN PEMBAHASAN

Hasil Pengumpulan dan Persiapan Data

Tahap awal penelitian melibatkan pengumpulan dan persiapan *dataset* sebelum dimasukkan ke dalam model pembelajaran mendalam. Tahapan ini sangat krusial untuk memastikan model memiliki generalisasi yang baik serta terhindar dari bias. Penelitian memanfaatkan tiga jenis dataset, yaitu *dataset real text manipulation* (RTM) sebagai fondasi pengetahuan sistem karena memiliki *mask* anotasi tingkat piksel, *dataset* media sosial (sosmed) yang memiliki karakteristik kompresi tingkat tinggi dan variasi format untuk mensimulasikan kondisi dunia nyata, serta *dataset AI* yang digunakan khusus untuk mengevaluasi kemampuan model menangani manipulasi modern tanpa artefak tepi kasar.

Tabel 2. Hasil Pengumpulan Dataset

Jenis Dataset	Fungsi Pengujian	Kelas Asli	Kelas Manipulasi	Total Citra
Real Text Manipulation (RTM)	Pelatihan, Validasi, dan Pengujian Dasar	3.000	6.000	9.000
Dataset Sosmed	Pengujian Generalisasi	100	100	200
Dataset AI	Pengujian Ketahanan Model	100	100	200
Total Keseluruhan		3.200	6.200	9.400

Pada observasi awal terhadap *dataset* utama (RTM), ditemukan adanya ketidakseimbangan kelas yang sangat signifikan, di mana kelas manipulasi mendominasi kelas asli dengan rasio 2:1. Untuk mencegah bias akurasi dan pergeseran sebaran (*covariate shift*), diterapkan prosedur re-stratifikasi *dataset* dengan menggabungkan seluruh data RTM beserta data suntikan, mengacaknya, lalu membaginya secara proporsional. Data minoritas juga digandakan melalui teknik *oversampling*. Selanjutnya, untuk mengoptimalkan penanganan ketidakseimbangan kelas pada tahap pelatihan, diimplementasikan mekanisme *dynamic class weighting* pada fungsi kerugian *binary cross entropy with logits loss* (BCEWithLogitsLoss). Berdasarkan log distribusi kelas (2.740 asli dan 4.719 manipulasi), koefisien bobot dinamis sebesar 0,5806 diinisialisasikan ke dalam *criterion loss*. Nilai yang bertindak sebagai faktor penekan ini terbukti efektif mencegah *overfitting* dengan memberikan penalti yang lebih berat jika model salah mengklasifikasikan kelas minoritas.

Implementasi Arsitektur Multi-stream Network

Arsitektur *multi-stream network* diimplementasikan menggunakan *framework* PyTorch dengan dua jalur ekstraksi utama. Pada jalur visual, citra masukan dilewatkan pada lapisan filter tepi kustom (Sobel) untuk menghasilkan *edge map* yang ditambahkan kembali ke citra asli dengan bobot intensitas 0,3 guna menajamkan area batas karakter. Citra yang telah dipertajam ini diekstraksi menggunakan *backbone* EfficientNet-B3 dengan *pre-trained weights* dari ImageNet, di mana lapisan pengklasifikasi aslinya diganti dengan lapisan *bottleneck* kustom untuk menghasilkan vektor fitur visual berdimensi 256. Secara bersamaan, jalur tekstual memproses citra menggunakan pustaka EasyOCR untuk mengekstrak lima parameter statistik (rata-rata *confidence score*, variansi posisi vertikal, rasio kepadatan, serta *confidence score* maksimum dan minimum). Tensor fitur mentah ini diekspansi melalui dua tahapan lapisan ber-dropout menjadi vektor fitur tekstual final berdimensi 128. Kedua vektor tersebut kemudian disatukan menggunakan operasi konkatenasi menjadi vektor gabungan berdimensi 384. Vektor ini diproses oleh lapisan *gated attention mechanism* berbasis aktivasi sigmoid untuk menghasilkan bobot gerbang dinamis yang dikalikan kembali melalui *element-wise multiplication*. Fitur terboboti yang hak suaranya telah diseimbangkan ini masuk ke blok *fusion classifier* akhir berdimensi 128 dengan *dropout* 50% untuk mengeluarkan probabilitas klasifikasi tunggal.

Proses Pelatihan Model

Pelatihan model dioptimalkan menggunakan algoritma AdamW dan dibagi menjadi dua tahap untuk menjaga stabilitas parameter. Tahap *warmup*

dilakukan dengan membekukan (*freezing*) parameter *backbone* EfficientNet-B3 dan melatih lapisan pengklasifikasi menggunakan *learning rate* 1e-3. Tahap *fine-tuning* selanjutnya membuka pembekuan (*unfreezing*) pada blok fitur ke-5 hingga ke-8, melatih model secara menyeluruh menggunakan *differential learning rate* (5e-5 untuk CNN dan 5e-4 untuk OCR/fusi), serta dikawal oleh *learning rate scheduler*. Mekanisme *best model checkpointing* juga diterapkan pada setiap akhir iterasi epoch untuk menyimpan bobot optimal secara otomatis saat nilai *validation loss* menyentuh titik terendah.

Hasil Pengujian dan Evaluasi Model

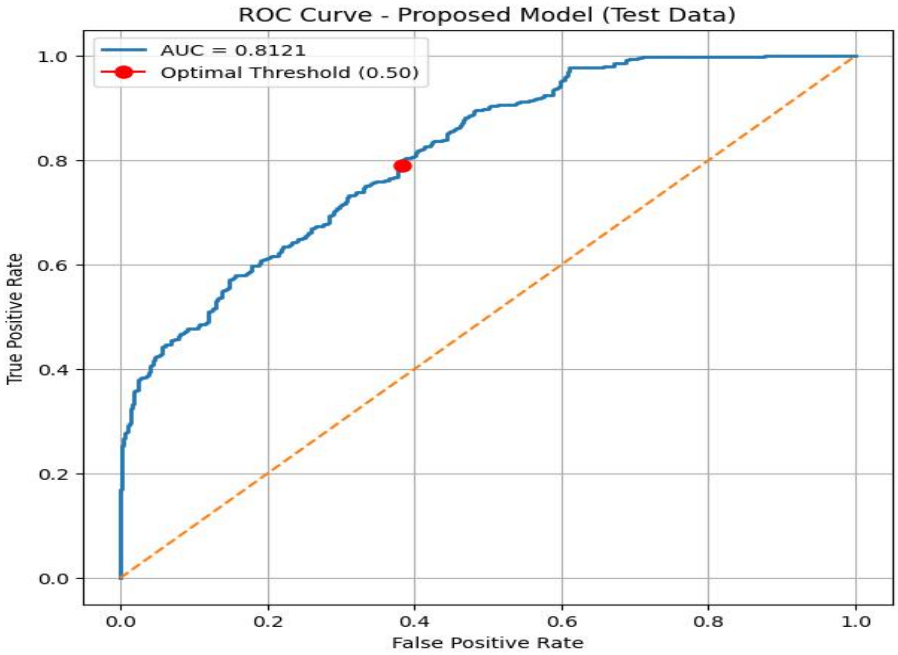
Keluaran akhir dari arsitektur *multi-stream network* adalah nilai probabilitas kontinu. Mengingat distribusi data yang tidak seimbang, ambang batas (*threshold*) dievaluasi melalui analisis kurva *precision-recall* menggunakan fungsi objektif argmax untuk memaksimalkan *F1-score*, dengan persamaan:

$$Threshold_{opt} = argmax_t \left(\frac{2 \times Precision(t) \times Recall(t)}{Precision(t) + Recall(t)} \right)$$

Berdasarkan perhitungan tersebut, *threshold* klasifikasi optimal bergeser menjadi 0,4961. Pengujian menghasilkan nilai *area under curve* (AUC) sebesar 0,8121. Nilai tersebut membuktikan bahwa model memiliki probabilitas keandalan sebesar 81,21% untuk secara tepat membedakan antara citra hasil manipulasi dan citra asli.

Titik merah pada kurva merepresentasikan *optimal threshold* sebesar 0,4961. Pada titik tersebut, model mencapai *true positive rate* (TPR) sekitar 0,79 dengan *false positive rate* (FPR) sekitar 0,38, yang menunjukkan kompromi antara kemampuan mendeteksi citra manipulasi dan tingkat kesalahan klasifikasi.

Nilai ambang batas 0,4961 inilah yang kemudian ditetapkan sebagai titik pisah absolut. Jika probabilitas keluaran model $\geq 0{,}4961$, maka citra diprediksi sebagai "Manipulasi". Sebaliknya, jika $< 0{,}4961$, citra diprediksi sebagai "Asli".



Gambar 2. Kurva ROC pada Pengujian Model Usulan

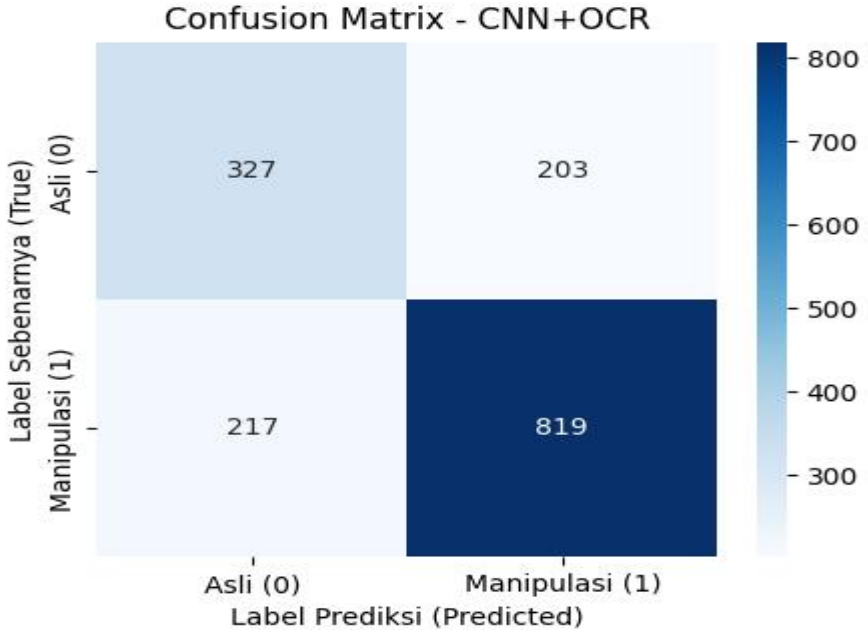
Pengujian *ablation study* dilakukan menggunakan 1.566 citra uji dari

dataset RTM untuk membandingkan empat skenario arsitektur.

Tabel 3. Hasil Pengujian *Ablation Study*

Skenario Pengujian	Accuracy	Precision	Recall	F1-Score
Skenario 1 (CNN)	0,6501	0,8398	0,5820	0,6876
Skenario 2 (CNN dan Sobel)	0,6775	0,8374	0,6361	0,7230
Skenario 3 (OCR)	0,6220	0,7688	0,6129	0,6821
Skenario 4 (CNN, Sobel, OCR, dan Attention)	0,7318	0,8014	0,7905	0,7959

Hasil *ablation study* menunjukkan bahwa metode usulan (skenario 4) yang menggabungkan fitur visual dan tekstual dengan mekanisme atensi berhasil mengungguli seluruh model *baseline*. Integrasi informasi tipografi OCR terbukti menambal kelemahan CNN pada area manipulasi yang sulit dibedakan. Model usulan mencapai akurasi sebesar 73,18% dengan nilai *F1-score* 0,7959. Meskipun terjadi sedikit penurunan pada metrik presisi (0,8014) akibat fenomena *precision-recall trade-off* dari efek penajaman tepi, hal ini dikompensasi dengan peningkatan *recall* menjadi 0,7905, membuktikan model menjadi jauh lebih peka dalam menangkap anomali.



Gambar 3. Confusion Matrix Skenario Metode Usulan

Berdasarkan visualisasi *confusion matrix*, model menunjukkan jumlah *true positive* (TP) dan *true negative* (TN) yang mendominasi, meski sebagian kecil citra masih mengalami *false positive* (FP) atau *false negative* (FN) akibat kemiripan karakteristik visual antarkelas yang ekstrem.

Analisis Signifikansi Statistik Menggunakan Uji McNemar

Penelitian ini menerapkan uji statistik McNemar dengan koreksi kontinuitas Edwards ($\alpha = 0{,}05$) untuk membandingkan prediksi dari dua skenario pada 1.566 citra uji yang sama. Hasil pengujian antara model unimodal baseline (CNN) dan model multimodal usulan (CNN dan OCR) menghasilkan nilai statistik Chi-Squared sebesar 39,3390 dengan P-Value sebesar $3{,}5625 \times 10^{-10}$. Nilai signifikansi yang jauh di bawah ambang batas 0,05 menunjukkan bahwa hipotesis nol (H_0) ditolak, sehingga terdapat perbedaan performa yang signifikan secara statistik antara kedua model. Selain itu, model usulan berhasil

mengklasifikasikan dengan benar 269 citra yang sebelumnya salah diklasifikasikan oleh CNN murni, sedangkan CNN murni hanya berhasil mengklasifikasikan dengan benar 141 citra yang salah diklasifikasikan oleh model usulan.

Perbandingan dengan Metode Lain pada Dataset RTM

Evaluasi terhadap *dataset* RTM menunjukkan keunggulan pendekatan multimodal yang diusulkan dibandingkan metode terdahulu. Sebagian besar penelitian *state-of-the-art* pada dataset ini, termasuk referensi dasar oleh pembuat dataset (Luo et. al), umumnya masih berfokus pada pengembangan fitur visual berbasis pergeseran distribusi RGB atau frekuensi. Metode unimodal tersebut sering kali mengalami penurunan *recall* ketika berhadapan dengan manipulasi teks dengan resolusi tinggi. Penelitian usulan memberikan *novelty* dengan membuktikan bahwa penambahan fitur statistik anomali tipografi dari OCR mampu menambal titik buta pendeteksian visual. Peningkatan *recall* dari 0,5820 (CNN murni) menjadi 0,7905 (Multimodal) menegaskan bahwa arsitektur fusi mampu mendeteksi anomali susunan teks meskipun batas piksel hasil manipulasi tersamarkan dengan sangat halus.

Analisis Kompleksitas Komputasi dan Waktu Inferensi

Penggunaan EfficientNet-B3 sebagai *backbone* visual secara inheren dirancang untuk meminimalkan kompleksitas waktu dan parameter komputasi, yang dibuktikan dengan waktu inferensi yang sangat cepat pada skenario CNN murni (66,89 ms). Sementara itu, integrasi pustaka EasyOCR memberikan beban komputasi tambahan yang signifikan karena proses pemindaian dan pengenalan karakter spasial di seluruh area citra membutuhkan sumber daya tinggi. Merujuk pada pengujian yang dilakukan pada 50 citra sampel acak, Tabel 4 menunjukkan perbandingan waktu inferensi rata-rata yang dieksekusi pada lingkungan *graphics processing unit* (GPU).

Tabel 4. Waktu Inferensi Model per Citra

Skenario Pengujian	Waktu Inferensi
Skenario 1 (CNN *Baseline*)	66,89 ms
Skenario 2 (CNN dan Sobel)	62,37 ms
Skenario 3 (OCR)	2084,20 ms
Skenario 4 (CNN dan OCR)	1787,70 ms

Dari data pada Tabel 4, terlihat bahwa ekstraksi OCR menjadi *bottleneck* utama dalam komputasi waktu. Walaupun waktu inferensi pada model multimodal usulan (1787,70 ms) jauh lebih lambat dibandingkan model CNN murni, rentang waktu di bawah 2 detik per citra masih sangat wajar dan praktis untuk skenario forensik digital.

Pengujian Generalisasi dan Ketahanan Model

Skenario pengujian generalisasi dan ketahanan dihadapkan pada 40 citra uji dari media sosial dan 40 citra manipulasi AI.

Tabel 5. Hasil Pengujian Generalisasi Media Sosial

Metrik Evaluasi	Nilai Persentase
Accuracy	50,00%
Precision	50,00%
Recall	25,00%

F1-Score	33,33%

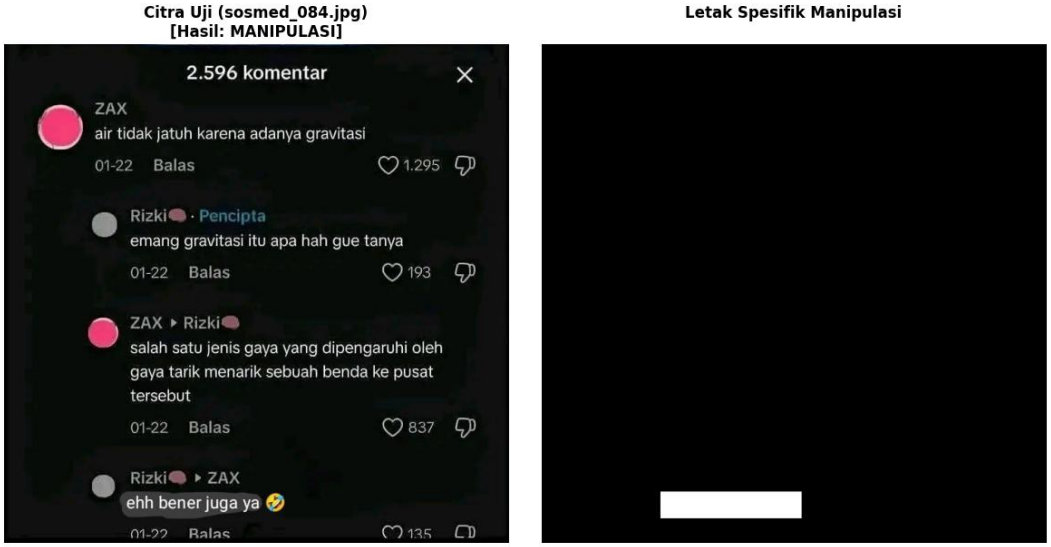
Tabel 6. Hasil Pengujian Ketahanan Terhadap AI

Metrik Evaluasi	Nilai Persentase
Accuracy	45,00%
Precision	44,44%
Recall	40,00%
F1-Score	42,11%

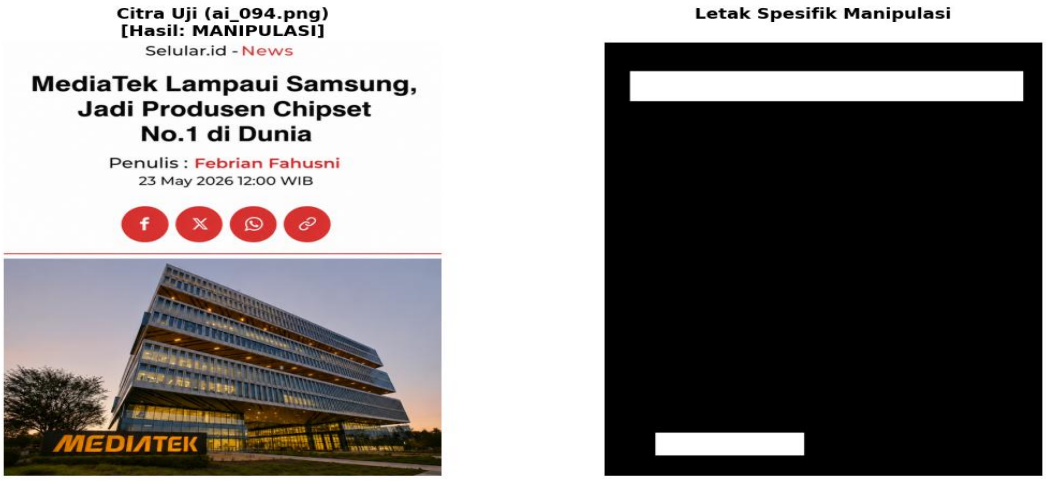
Pada uji media sosial, terjadi penurunan kinerja yang signifikan (Akurasi 50,00%, *Recall* 25,00%), mengonfirmasi adanya *domain gap*. Kompresi JPEG yang agresif menghasilkan *noise* kompresi yang menyebar di seluruh piksel gambar, mengacaukan *edge map* dari filter Sobel sehingga model kesulitan membedakan antara artefak manipulasi dan artefak kompresi bawaan. Uji ketahanan terhadap manipulasi AI menunjukkan tantangan yang lebih berat. Diskrepansi piksel lokal berhasil disamarkan secara sempurna oleh algoritma generatif *context-aware fill*. Walaupun OCR mencoba menangkap anomali tata letak, kemampuan adaptasi masih terbatas akibat dominasi data manipulasi manual (RTM) saat pelatihan.

Analisis Kualitatif

Uji analisis kualitatif dilakukan dengan mengumpankan citra spesifik ke dalam sistem dan melacak letak *bounding box*.



Gambar 4. Citra Sosmed yang Berhasil Dideteksi



Gambar 5. Citra AI yang Berhasil Dideteksi

Gambar 6. Citra Sosmed yang Gagal Dideteksi

Gambar 7. Citra AI yang Gagal Dideteksi

Pada sampel yang berhasil (TP) seperti *sosmed_084.jpg* dan *ai_094.png*, probabilitas klasifikasi melampaui 0,4961 dan OCR secara akurat memetakan kotak pembatas tepat pada teks anomali. Sebaliknya, pada kasus kegagalan deteksi (FN) seperti *sosmed_088.jpg* (0,4590) dan *ai_099.png* (0,4484), kesalahan identifikasi dipicu oleh kehalusan *blending* algoritma *generative AI* yang tidak memicu aktivasi batas kontras pada CNN, serta konsistensi teks palsu yang sangat bersih sehingga pustaka EasyOCR memberikan nilai *confidence score* yang tinggi menyerupai teks asli.

KESIMPULAN

Berdasarkan hasil evaluasi penelitian, dapat ditarik beberapa poin kesimpulan utama. Pertama, model deteksi manipulasi timpa gambar dengan teks berhasil dibangun melalui integrasi *convolutional neural network* (EfficientNet-B3) dan *optical character recognition* (EasyOCR) ke dalam arsitektur *multi-stream network*. Model sukses dilatih menggunakan mekanisme *gated attention*, *dynamic class weighting* sebesar 0,5806, serta pendekatan *differential learning rate* untuk mengatasi ketidakseimbangan kelas. Kedua, performa deteksi model usulan terbukti lebih unggul dengan mencapai akurasi 73,18% dan *F1-score* 0,7959 pada *threshold* 0,4961, melampaui model CNN visual tunggal. Integrasi informasi tekstual OCR menambal kelemahan CNN pada area artefak halus yang tersamarkan, ditandai dengan pencapaian tingkat *recall* sebesar 79,05%. Ketiga, performa generalisasi model mengalami penurunan signifikan pada pengujian *dataset* media sosial (akurasi 50,00%, *F1-score* 33,33%) dan *dataset* manipulasi AI (akurasi 45,00%, *F1-score* 42,11%), yang mengonfirmasi adanya *domain gap* akibat kesulitan model beradaptasi dengan *compression noise* ekstrem serta kehalusan algoritma *generative AI*.

Sebagai penelitian lanjutan, keterbatasan generalisasi pada *unseen domain* dapat diatasi dengan memperkaya variasi *dataset* pelatihan secara masif (*advanced data augmentation*) dari citra media sosial, serta menerapkan teknik *domain adaptation* untuk membiasakan model terhadap distribusi *noise* dunia nyata. Selain itu, untuk melengkapi sistem deteksi agar tahan terhadap *context-aware blending AI*, pengembangan selanjutnya sangat direkomendasikan untuk menambahkan ekstraksi fitur berbasis domain frekuensi (seperti DCT atau SRM) serta mengintegrasikan *natural language processing* (NLP) untuk menganalisis inkonsistensi makna linguistik. Terakhir, eksplorasi menggunakan arsitektur *transformer-based multimodal model* juga berpotensi besar untuk diimplementasikan di masa depan guna menangkap korelasi spasial global

antara anomali teks dan gambar secara lebih presisi.

DAFTAR PUSTAKA

[1] Y. Yang, T. Davis, and M. Hindman, "Visual misinformation on Facebook," *J. Commun.*, no. December 2022, pp. 316–328, 2023.

[2] L. S. Putra, Mahsun, A. Sirulhaq, and Saharudin, "Tindak pidana verbal dunia maya (cyber crime) pada media sosial Facebook: Kajian linguistik forensik," *J. Kreat. Pendidik. Mod.*, vol. 7, no. 3, pp. 59–81, 2025.

[3] Komdigi, "Penanganan Konten Hoaks Tahun 2023," *J. Stud. Komun. dan Media*, vol. 28, no. 2, pp. 167–182, 2024.

[4] Komdigi, "Kominfo Klarifikasi 1.923 Konten Hoaks Sepanjang Tahun 2024." Accessed: May 06, 2026. [Online]. Available: https://www.komdigi.go.id/berita/siaran-pers/detail/komdigi-identifikasi-1923-konten-hoaks-sepanjang-tahun-2024

[5] M. A. I. (MAFINDO), "Laporan Pemetaan Hoaks 2023," 2024. [Online]. Available: https://www.mafindo.or.id/

[6] M. Zhu, M. Li, and Z. Wang, "Image tampering detection based on RDS-YOLOv5 feature enhancement transformation," *Pattern Recognit.*, pp. 1–15, 2024.

[7] R. Jadhav, V. Meshram, A. Bhosle, K. Patil, S. Dash, and S. Jadhav, "Explainable multilingual and multimodal fake-news detection: toward robust and trustworthy AI for combating misinformation," *Front. Artif. Intell.*, no. December, pp. 1–20, 2025, doi: 10.3389/frai.2025.1690616.

[8] Y.-Y. Bae, D.-J. Cho, and K.-H. Jung, "Enhancing Document Forgery Detection with Edge-Focused Deep Learning," *Symmetry (Basel).*, 2025, [Online]. Available: https://doi.org/10.3390/sym17081208

[9] L. Wang, Z. Li, and W. Zhao, "Research on identity document image tampering detection based on texture understanding and multistream networks," *J. Electron. Imaging*, vol. 34, no. 4, pp. 1–21, 2025, doi: 10.1117/1.JEI.34.4.043018.

[10] D. Luo *et al.*, "Toward real text manipulation detection: New dataset and new solution," *Pattern Recognit.*, vol. 157, p. 110828, 2025, doi: https://doi.org/10.1016/j.patcog.2024.110828.

[11] S. El-amrany, S. Lamsiyah, M. R. Brust, and P. Bouvry, "GuardHarMem and HarMDetect: a multimodal dataset and benchmark model for fine-grained harmful meme classification," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, pp. 1–14, 2025, doi: 10.1007/s13278-025-01475-2.

[12] C. Dong, X. Chen, R. Hu, J. Cao, and X. Li, "MVSS-Net: Multi-View Multi-Scale Supervised Networks for Image Manipulation Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3539–3553, 2023, doi: 10.1109/TPAMI.2022.3180556.

[13] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 20606–20615.

[14] D. V. Bui, T. D. Tran, H. D. Tran, V. K. Dinh, D. C. Pham, and T. L. Dang, "A lightweight dual-branch spatial-frequency framework for joint deepfake detection and localization," *Comput. Vis. Image Underst.*, vol. 270, p. 104860, 2026, doi: https://doi.org/10.1016/j.cviu.2026.104860.