Indonesian Journal of Multidisciplinary on Social and Technology
Homepage: https://journal.ilmudata.co.id/index.php/ijmst
Vol. 5 No. 1 (2026) pp: 1153-1162
P-ISSN: 2986-6790, e-ISSN: 2986-6782

Model ConvNext Berbasis Fusi Fitur Spektral untuk Deteksi Citra Sintetis yang Robust

Reyvan Revolusioner Ar, Adha Mashur Sajiah, Sutardi
Jurusan Informatika, Fakultas Teknik, Universitas Halu Oleo, Kendari, Indonesia
reyvanrevolusionerar12@gmail.com, adha.m.sajiah@uho.ac.id, sutardi@uho.ac.id

Abstrak

Model generatif semakin mampu menghasilkan citra realistis sehingga deteksi citra sintetis menjadi sulit setelah citra mengalami kompresi, pengaburan, derau, atau perubahan ukuran. Penelitian ini mengembangkan detektor ConvNeXt-Tiny yang menggabungkan fitur spasial dengan lima representasi forensik: magnitudo Fourier, discrete cosine transform (DCT), wavelet Haar, fase Fourier, dan residual spatial rich model (SRM). Keenam cabang diproyeksikan menjadi embedding 256 dimensi dan diintegrasikan melalui gated residual fusion untuk menghasilkan kepala klasifikasi spatial, spectral, dan fusion. Evaluasi eksperimen mencakup klasifikasi bersih, 21 kondisi pascapemrosesan, adaptasi domain baru, regresi domain lama, dan pengujian white box. Pada benchmark model dasar, fusion memperoleh akurasi 93,00% pada citra bersih dan mengungguli spatial dalam 18 dari 22 kondisi. Keunggulan terbesar sebesar 1,69 poin persentase terjadi pada resize 0,05, sedangkan pada noise Gaussian sigma 0,10 fusion lebih rendah 1,50 poin persentase. Fine-tuning memilih epoch ketujuh dengan skor validasi gabungan 0,7991. Pada 263 citra uji domain baru yang dinormalisasi, akurasi fusion meningkat dari 53,23% menjadi 84,79% dan macro-F1 dari 0,4416 menjadi 0,8473. Uji McNemar exact menunjukkan perubahan sebelum dan sesudah fine-tuning yang signifikan ($p < 0,001$). Namun, akurasi pada 2.000 citra domain lama menurun dari 84,10% menjadi 80,55%. Sebelas kasus white box lulus, termasuk empat jalur dasar fungsi fusi. Hasil menunjukkan bahwa manfaat fusi spektral bergantung pada kondisi gangguan dan adaptasi domain menimbulkan kompromi antardomain. Ablation terkontrol dan perbandingan berpasangan antarkepala masih diperlukan untuk memastikan kontribusi cabang serta keunggulan statistik fusion terhadap spatial.

Kata kunci: ConvNeXt, Deteksi Citra Sintetis, Fusi Fitur Spektral, Robustness, Adaptasi Domain

Abstract

Generative image models increasingly produce realistic images, making synthetic-image detection difficult after compression, blurring, noise, or resizing. This study develops a ConvNeXt-Tiny detector that combines spatial features with five forensic representations: Fourier-transform magnitude, discrete cosine transform (DCT), Haar wavelets, Fourier phase, and spatial rich model (SRM) residuals. The six branches are projected into 256-dimensional embeddings and integrated through gated residual fusion to produce spatial, spectral, and fusion classification heads. The experimental evaluation covers clean classification, 21 post-processing conditions, adaptation to a new domain, regression on an earlier domain, and white-box testing. In the base-model benchmark, fusion achieved 93.00% accuracy on clean images and exceeded spatial in 18 of 22 conditions. Its largest advantage was 1.69 percentage points at resize factor 0.05, whereas it was 1.50 points lower under Gaussian noise with sigma 0.10. Fine-tuning selected epoch seven with a combined validation score of 0.7991. On 263 normalized new-domain test images, fusion accuracy increased from 53.23% to 84.79%, and macro-F1 from 0.4416 to 0.8473. An exact McNemar test indicated a significant before-and-after change ($p < 0.001$). However, accuracy on 2,000 earlier-domain images declined from 84.10% to 80.55%. All eleven white-box test cases passed, including four basis paths of the fusion function. The findings show condition-dependent benefits from spectral fusion and a measurable domain-adaptation trade-off. Controlled ablation and paired comparisons between classification heads remain necessary to establish individual branch contributions and statistical superiority over the spatial head.

Keywords: ConvNeXt, Synthetic Image Detection, Spectral Feature Fusion, Robustness, Domain Adaptation

1. Pendahuluan

Perkembangan model generatif telah meningkatkan kemampuan menghasilkan citra yang semakin realistis. Citra sintetis yang dihasilkan model difusi maupun generative adversarial networks (GAN) dapat menyerupai foto sehingga pemeriksaan visual tidak selalu cukup untuk memastikan keasliannya. Tantangan tersebut semakin besar ketika citra diproses ulang melalui platform digital. Kompresi JPEG dan perubahan resolusi dapat mengubah

petunjuk forensik yang sebelumnya digunakan detektor. Corvi et al. (2023) menunjukkan bahwa detektor yang dikembangkan untuk citra GAN tidak selalu dapat mengenali citra dari diffusion models, terutama pada skenario kompresi dan resize. Ricker et al. (2024) juga melaporkan penurunan kemampuan detektor GAN ketika diterapkan pada citra yang dihasilkan model difusi.

Salah satu pendekatan untuk mengenali citra sintetis adalah analisis domain frekuensi. Jejak generatif dapat muncul sebagai pola pada spektrum frekuensi, koefisien transformasi, maupun residual citra. Penelitian UGAD memanfaatkan frequency fingerprints untuk mendeteksi citra generatif lintas model (Alam et al., 2024), sedangkan Wißmann et al. (2024) menunjukkan bahwa discrete cosine transform dan power spectral density dapat dimanfaatkan untuk deteksi serta atribusi model. Analisis kompresi juga menemukan jejak pada domain spasial dan frekuensi, tetapi generalisasi artefak tersebut tetap menjadi tantangan (Bergmann et al., 2024). Temuan lain menunjukkan bahwa penambahan fitur frekuensi tidak selalu meningkatkan akurasi dibandingkan CNN konvensional (Nourji et al., 2025). Dengan demikian, efektivitas fitur spektral tidak dapat diasumsikan berlaku sama pada setiap arsitektur dan kondisi citra.

Pada sisi arsitektur, ConvNeXt merupakan ConvNet modern yang mengadopsi sejumlah prinsip desain jaringan visual kontemporer dan menunjukkan kinerja kompetitif pada tugas pengenalan citra (Liu et al., 2022). Pengembangan ConvNeXt V2 memperkuat kerangka tersebut melalui co-design arsitektur dan pembelajaran representasi (Woo et al., 2023). Pendekatan multi-attentional juga menekankan pentingnya petunjuk tekstur lokal pada deteksi manipulasi wajah (Zhao et al., 2021). Sejumlah penelitian deteksi deepfake juga mengembangkan pendekatan berbasis ConvNeXt, baik melalui pembelajaran prototipe maupun kombinasi dengan transformer (Ilyas et al., 2024; Amarasinghe & Arachchi, 2025). Model ensemble yang menggabungkan representasi spasial dan frekuensi turut menunjukkan potensi untuk menghadapi variasi data (Bhattacharjee et al., 2026). Namun, hasil penelitian terbaru tentang bias frekuensi menegaskan bahwa ketergantungan berlebihan pada pola frekuensi dapat berkaitan dengan persoalan robustness dan generalisasi (Liu et al., 2026).

Penelitian di Indonesia juga mulai menguji pendekatan CNN untuk mendeteksi citra manipulatif atau sintetis. Taufiqurrahman et al. (2025) melaporkan peningkatan performa melalui dilated convolution, sedangkan Hardiyanto dan Akbar (2025) menggabungkan fitur tekstur GLCM dengan CNN untuk klasifikasi citra hasil model difusi. Raharjo et al. (2025) mengkaji inkonsistensi tekstur dan pola pada citra wajah. Penelitian-penelitian tersebut memperlihatkan perkembangan pendekatan visual dan hybrid, tetapi masih diperlukan evaluasi yang secara eksplisit membandingkan representasi spasial dan spektral pada beberapa jenis gangguan pascapemrosesan serta mengamati konsekuensi adaptasi pada domain sebelumnya. Detektor yang hanya dinilai pada citra bersih berisiko menghasilkan estimasi performa yang terlalu optimistis untuk penggunaan pada citra yang telah mengalami pemrosesan ulang.

Berdasarkan celah tersebut, penelitian ini mengembangkan model enam cabang yang mengintegrasikan ConvNeXt-Tiny dengan magnitudo FFT, DCT, wavelet Haar, fase FFT, dan residual SRM. Setiap cabang diproyeksikan ke ruang embedding berdimensi sama dan digabungkan melalui gated residual fusion. Model menyediakan tiga kepala klasifikasi agar representasi spasial, spektral, dan gabungan dapat dievaluasi pada kerangka yang sama. Kontribusi penelitian terletak pada integrasi lima representasi forensik ke dalam arsitektur ConvNeXt, evaluasi robustness pada 22 kondisi citra, pengukuran adaptasi domain baru dan regresi domain lama, serta pemeriksaan jalur internal menggunakan white-box testing. Penelitian ini bertujuan mengukur performa fusi dibandingkan keluaran spasial pada kondisi pengujian yang tersedia, bukan mengklaim keunggulan statistik tanpa pengujian berpasangan antarkepala.

2. Metode Penelitian

2.1 Rancangan Penelitian dan Pengembangan Sistem

Penelitian menggunakan pendekatan eksperimen komputasional dalam pengembangan sistem deteksi citra sintetis. Pengembangan perangkat lunak mengikuti tahapan Rational Unified Process (RUP), yaitu inception, elaboration, construction, dan transition. Inception digunakan untuk menetapkan kebutuhan deteksi dan masalah robustness. Elaboration mencakup perancangan prapemrosesan, ekstraksi fitur, fusi, pelatihan, dan kalibrasi. Construction merealisasikan model ConvNeXt-Tiny dengan lima cabang forensik dan tiga kepala klasifikasi. Transition mencakup evaluasi citra bersih, gangguan pascapemrosesan, regresi domain lama, serta pengujian white box.

DOI: https://doi.org/10.69693/ijmst.v5i1.14703
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Implementasi menggunakan Python dan PyTorch; pengelolaan eksperimen serta evaluasi dijalankan melalui notebook pada lingkungan Kaggle.

2.2 Data Penelitian

Data dikelola melalui manifest eksperimen yang mencatat label, sumber, dan metadata citra. Kelas REAL diberi label 0, sedangkan AI/FAKE diberi label 1. Sumber wai_shutterstock_deepmedia dan ai_detection_1024 digunakan untuk pengembangan model, sementara ai_recognition_external digunakan sebagai external reference. Tahap fine-tuning menambahkan data domain baru dari folder dataku dan dataku_uji, serta menggunakan replay dari data domain sebelumnya. Komposisi subset yang digunakan dalam evaluasi dirangkum pada Tabel 1. Jumlah train gabungan yang tercatat adalah 7.200 entri, terdiri atas 1.800 entri domain baru dan 5.400 entri replay; jumlah tersebut mengikuti metadata checkpoint pendamping.

Tabel 1. Komposisi data penelitian

Subset	Jumlah	Komposisi/keterangan
Train domain baru	1.800	Data untuk adaptasi domain
Replay train	5.400	Data dari domain sebelumnya
Validation baru	141	71 REAL dan 70 AI
Validation replay	1.500	Data validasi domain sebelumnya
Test baru	263	133 REAL dan 130 AI
External reference	2.000	1.000 REAL dan 1.000 AI

Data validasi domain baru terdiri atas 141 citra, sedangkan data uji domain baru berjumlah 263 citra, yaitu 133 REAL dan 130 AI. External reference berisi 2.000 citra dengan komposisi seimbang, masing-masing 1.000 citra REAL dan AI. Citra uji domain baru dievaluasi dalam versi yang dinormalisasi dan versi asli; keduanya merupakan representasi dari 263 citra yang sama, bukan dua kelompok sampel independen. Pembagian data dilakukan sebelum pelatihan, dan data uji tidak digunakan untuk memilih epoch maupun threshold.

2.3 Prapemrosesan dan Ekstraksi Fitur

Citra dibaca dengan koreksi orientasi EXIF dan dikonversi ke RGB. Pada tahap training, satu area crop digunakan bersama agar jalur spasial dan forensik memproses wilayah citra yang sama. Masukan jalur spasial diubah menjadi 224 × 224 piksel, sedangkan masukan jalur forensik menjadi 128 × 128 piksel. Normalisasi ImageNet diterapkan hanya pada jalur spasial dengan mean kanal (0,485; 0,456; 0,406) dan standar deviasi (0,229; 0,224; 0,225). Jalur forensik menggunakan grayscale dari RGB sebelum normalisasi ImageNet. Dalam resep fine-tuning, sisi terpanjang citra domain baru dibatasi hingga 640 piksel dan kompresi JPEG berkualitas 65–90 ditentukan secara deterministik berdasarkan hash. Augmentasi training meliputi horizontal flip, perubahan brightness dan contrast, serta JPEG, blur, noise, dan resize yang diterapkan secara simetris pada kedua kelas.

Cabang spasial menggunakan ConvNeXt-Tiny tanpa classifier bawaan. Global average pooling menghasilkan 768 fitur yang diproyeksikan menjadi embedding 256 dimensi. Lima cabang forensik diproses dengan encoder konvolusional tersendiri, menghasilkan 192 fitur sebelum diproyeksikan menjadi 256 dimensi. Magnitudo FFT dihitung dari grayscale setelah pengurangan rata-rata, transformasi Fourier dua dimensi, dan pemusatan frekuensi melalui fftshift. Magnitudo kompleks diubah dengan $\log(1 + |F|)$ dan distandardisasi. Cabang fase menggunakan sudut kompleks dari transformasi Fourier. Cabang DCT menggunakan grayscale berukuran 64 × 64, koefisien DC dihilangkan, kemudian diterapkan transformasi log-magnitudo dan standardisasi.

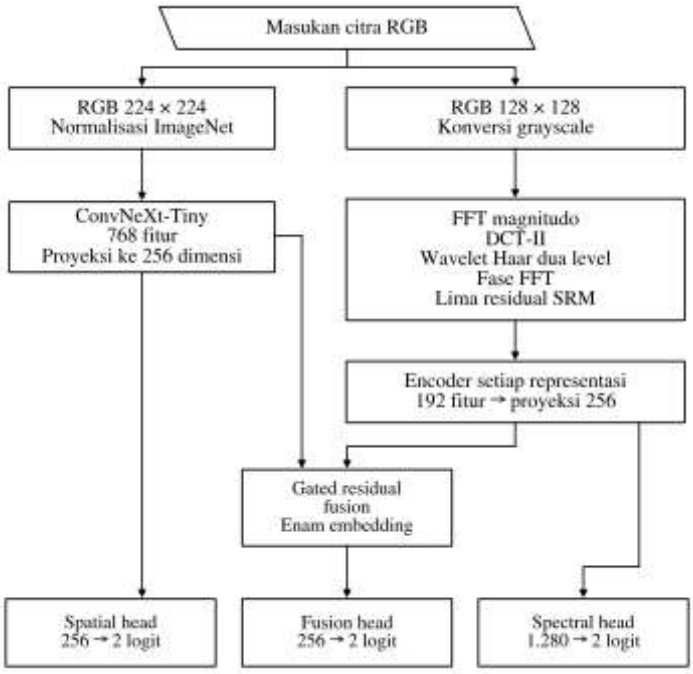
Cabang wavelet Haar menghasilkan koefisien LL, LH, HL, dan HH pada level pertama. Dekomposisi level kedua diterapkan pada LL level pertama; hasilnya disesuaikan dan digabungkan menjadi delapan kanal berukuran 64 × 64. Cabang SRM menerapkan lima kernel residual tetap, kemudian menggunakan $\log(1 + |r|)$ dan standardisasi dengan epsilon 10^{-6} untuk menjaga kestabilan numerik. Dimensi masukan dan embedding keenam cabang

DOI: https://doi.org/10.69693/ijmst.v5i1.14703
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

ditampilkan pada Tabel 2. Arsitektur keseluruhan dan hubungan antara jalur spasial, jalur forensik, fusi, serta kepala klasifikasi ditunjukkan pada Gambar 1.

Tabel 2. Dimensi masukan dan embedding enam cabang

Cabang	Representasi masukan	Fitur encoder	Embedding
Spatial	3 × 224 × 224	768	256
FFT	1 × 128 × 128	192	256
DCT	1 × 64 × 64	192	256
Wavelet Haar	8 × 64 × 64	192	256
Fase FFT	1 × 128 × 128	192	256
SRM	5 × 128 × 128	192	256



Gambar 1. Arsitektur enam cabang dan tiga kepala klasifikasi

2.4 Fusi Fitur, Pelatihan, dan Kalibrasi

Enam embedding disusun menjadi tensor berukuran B × 6 × 256, dengan B menyatakan ukuran batch. Modul gated residual fusion menghitung konteks dari rata-rata embedding yang dipertahankan, kemudian menghasilkan bobot gate melalui softmax. Embedding berbobot dijumlahkan dan dikombinasikan dengan jalur residual berbasis multilayer perceptron, lalu dinormalisasi. Pada training, branch dropout dapat menonaktifkan sebagian cabang. Jika seluruh cabang suatu sampel terhapus, cabang spatial dipulihkan. Pada evaluasi, keenam cabang digunakan. Spatial head menerima embedding spasial 256 dimensi, spectral head menerima concatenation lima embedding forensik berukuran 1.280, sedangkan fusion head menerima embedding hasil fusi 256 dimensi. Setiap head menghasilkan dua logit untuk kelas REAL dan AI.

Fine-tuning dimulai dari checkpoint model dasar. Tahap ft_heads memperbarui kepala klasifikasi, kemudian tahap ft_core membuka encoder forensik, proyeksi, modul fusi, dan stage terakhir backbone. Optimisasi menggunakan AdamW dengan weight decay 0,05. Base learning rate adalah $2{,}5 \times 10^{-4}$; pada ft_heads learning rate head sebesar $1{,}25 \times 10^{-4}$, sedangkan pada ft_core learning rate head, core, dan backbone masing-masing $2{,}5 \times 10^{-5}$, $7{,}5 \times 10^{-6}$, dan $1{,}25 \times 10^{-5}$. Tahap ft_heads berlangsung maksimal tiga epoch dengan patience dua, sedangkan ft_core maksimal 12 epoch dengan patience empat. Fungsi loss menggabungkan cross-entropy spatial, spectral, dan fusion dengan bobot 0,30; 0,30; dan 0,40, serta label smoothing 0,05. Pelatihan dievaluasi pada validation baru dan

DOI: https://doi.org/10.69693/ijmst.v5i1.14703
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

validation replay. Skor gabungan merupakan rata-rata skor macro-F1 kedua domain. Checkpoint dipilih jika batas performa domain lama terpenuhi dan skor gabungan meningkat. Epoch terbaik dikalibrasi menggunakan temperature scaling; threshold dicari pada rentang 0,30–0,70 dengan langkah 0,005 berdasarkan macro-F1 validation, dan nilai yang lebih dekat ke 0,5 diprioritaskan ketika skor seri.

2.5 Skenario dan Metrik Evaluasi

Benchmark robustness model dasar membandingkan spatial dan fusion pada satu kondisi bersih serta 21 kondisi gangguan. Gangguan terdiri atas enam tingkat kualitas JPEG (90, 70, 50, 30, 20, dan 10), lima radius Gaussian blur (0,5; 1; 2; 3; dan 4), lima tingkat Gaussian noise (sigma 0,01; 0,02; 0,04; 0,06; dan 0,10), serta lima faktor resize (0,75; 0,50; 0,25; 0,10; dan 0,05). Keunggulan fusion dihitung sebagai selisih akurasi fusion dan spatial pada kondisi yang sama. Rata-rata kondisi gangguan menggunakan bobot sama untuk setiap kondisi.

Evaluasi domain baru menggunakan akurasi, macro-F1, false positive rate (FPR), false negative rate (FNR), dan area under the ROC curve (AUC). Kelas positif adalah AI. FPR menggambarkan proporsi citra REAL yang keliru diprediksi AI, sedangkan FNR menggambarkan proporsi citra AI yang keliru diprediksi REAL. Perubahan sebelum dan sesudah fine-tuning pada 263 citra yang sama diuji menggunakan McNemar exact. Pengujian ini menilai perubahan prediksi model fusion sebelum dan sesudah fine-tuning, bukan signifikansi perbedaan fusion terhadap spatial. Regresi domain lama diukur pada 2.000 citra external reference.

2.6 Pengujian White Box

Pengujian white box memeriksa alur kontrol fungsi GatedResidualFusion.forward(), transformasi fitur, dimensi tensor, pemisahan masukan head, aliran gradien, pencarian threshold, dan penggabungan prediksi kalibrasi. Fungsi fusi memiliki empat jalur dasar yang mencakup mode evaluasi, training tanpa dropout, training dengan cabang yang bertahan, serta pemulihan cabang spatial ketika seluruh cabang terhapus. Pengujian dilaksanakan pada 29 Oktober 2026 menggunakan CPU, Python 3.12.14, PyTorch 2.15.0+cpu, seed 42, dan model tanpa bobot pralatih. Sebanyak sebelas kasus uji dijalankan untuk memeriksa kesesuaian hasil aktual dengan keluaran yang diharapkan.

3. Hasil dan Pembahasan

3.1 Implementasi Arsitektur dan Komposisi Data

Implementasi menghasilkan enam cabang ekstraksi fitur dan tiga kepala klasifikasi. Penyamaan embedding ke dimensi 256 memungkinkan fitur ConvNeXt-Tiny, magnitudo FFT, DCT, wavelet Haar, fase FFT, dan residual SRM diproses oleh satu modul gated residual fusion. Keberadaan spatial head menyediakan pembanding yang menggunakan representasi ConvNeXt, spectral head menggabungkan lima representasi forensik, dan fusion head menggunakan hasil integrasi keduanya. Perlu diperhatikan bahwa spatial head berada dalam model multihead dengan encoder yang digunakan bersama selama optimisasi; pembanding ini bukan model ConvNeXt independen yang dilatih terpisah.

Komposisi data menunjukkan bahwa evaluasi dilakukan pada subset domain baru dan domain lama. Penggunaan replay sebanyak 5.400 entri dimaksudkan untuk mempertahankan informasi dari domain sebelumnya ketika model beradaptasi pada 1.800 entri baru. Data uji baru berjumlah 263 citra dengan distribusi kelas yang hampir seimbang, sedangkan external reference berisi 2.000 citra seimbang. Pemisahan peran data tersebut penting karena hasil benchmark robustness model dasar, hasil fine-tuning pada domain baru, dan hasil regresi domain lama menjawab pertanyaan evaluasi yang berbeda.

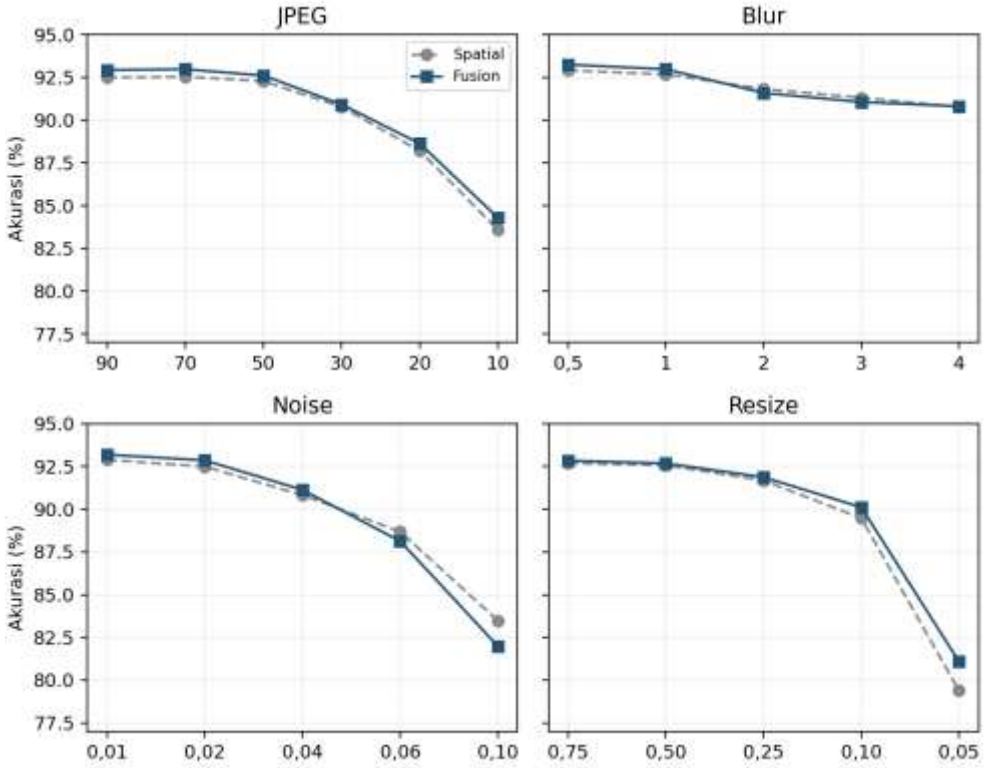
3.2 Kinerja Model Dasar pada Gangguan Pascapemrosesan

Pada kondisi bersih, fusion memperoleh akurasi 93,00%, sedikit lebih tinggi daripada spatial sebesar 92,70%. Secara keseluruhan, fusion mengungguli spatial pada 18 dari 22 kondisi pengujian. Pada 21 kondisi gangguan, rata-rata akurasi spatial adalah 90,17%, sedangkan fusion mencapai 90,38%, sehingga selisih rata-ratanya sekitar 0,21 poin persentase. Nilai ini menunjukkan peningkatan numerik yang terbatas pada rata-rata seluruh gangguan, meskipun terdapat kondisi tertentu dengan keuntungan yang lebih besar. Rincian akurasi pada setiap kondisi ditampilkan pada Tabel 3, sedangkan pola perubahan untuk JPEG, blur, noise, dan resize disajikan pada Gambar 2.

DOI: https://doi.org/10.69693/ijmst.v5i1.14703
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Tabel 3. Akurasi spatial dan fusion pada 22 kondisi pengujian

Kondisi	Spatial (%)	Fusion (%)	Δ (pp)
Bersih	92,70	93,00	+0,30
JPEG q90	92,49	92,93	+0,44
JPEG q70	92,54	92,98	+0,44
JPEG q50	92,28	92,61	+0,33
JPEG q30	90,78	90,92	+0,14
JPEG q20	88,20	88,63	+0,43
JPEG q10	83,58	84,29	+0,71
Blur 0,5	92,91	93,25	+0,34
Blur 1	92,68	92,98	+0,30
Blur 2	91,82	91,59	−0,23
Blur 3	91,31	91,06	−0,25
Blur 4	90,78	90,81	+0,03
Noise 0,01	92,89	93,19	+0,30
Noise 0,02	92,49	92,86	+0,37
Noise 0,04	90,83	91,13	+0,30
Noise 0,06	88,68	88,13	−0,55
Noise 0,10	83,48	81,98	−1,50
Resize 0,75	92,70	92,84	+0,14
Resize 0,50	92,56	92,68	+0,12
Resize 0,25	91,68	91,87	+0,19
Resize 0,10	89,47	90,11	+0,64
Resize 0,05	79,39	81,08	+1,69



Gambar 2. Perbandingan akurasi spatial dan fusion pada gangguan pascapemrosesan

DOI: https://doi.org/10.69693/ijmst.v5i1.14703
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Pada seluruh tingkat kompresi JPEG, fusion memperoleh akurasi lebih tinggi daripada spatial. Pada kualitas JPEG 10, fusion mencapai 84,29% dibandingkan 83,58% pada spatial, dengan selisih 0,71 poin persentase. Keunggulan juga terlihat pada seluruh faktor resize; pada faktor 0,05, fusion mencapai 81,08% sedangkan spatial 79,39%, sehingga menghasilkan selisih terbesar sebesar 1,69 poin persentase. Temuan ini mengindikasikan bahwa representasi gabungan memberikan informasi pelengkap ketika detail citra berkurang akibat kompresi atau penurunan resolusi.

Keunggulan tersebut tidak berlaku pada semua gangguan. Pada blur dengan radius 2 dan 3, akurasi fusion lebih rendah masing-masing 0,23 dan 0,25 poin persentase. Pada noise sigma 0,06 dan 0,10, fusion juga lebih rendah 0,55 dan 1,50 poin persentase. Selisih negatif terbesar pada noise 0,10 menunjukkan bahwa penambahan cabang forensik tidak selalu meningkatkan ketahanan. Pola ini sejalan dengan temuan bahwa representasi frekuensi dapat sensitif terhadap distribusi gangguan dan bias tertentu (Bergmann et al., 2024; Liu et al., 2026). Karena itu, hasil benchmark lebih tepat ditafsirkan sebagai manfaat fusi yang bergantung pada jenis dan intensitas gangguan, bukan keunggulan universal.

Frekuensi keunggulan fusion pada konfigurasi enam cabang juga lebih tinggi daripada konfigurasi pembanding tujuh cabang yang menambahkan representasi radial. Konfigurasi enam cabang unggul pada 18 dari 22 kondisi, sedangkan konfigurasi tujuh cabang unggul pada 12 dari 22 kondisi. Perbandingan ini menunjukkan bahwa penambahan satu representasi tidak otomatis memperluas ketahanan model. Namun, hasil tersebut belum menjelaskan kontribusi kausal setiap cabang karena penelitian belum melakukan ablation terkontrol dengan penghilangan satu cabang secara sistematis.

3.3 Hasil Fine-Tuning pada Domain Baru

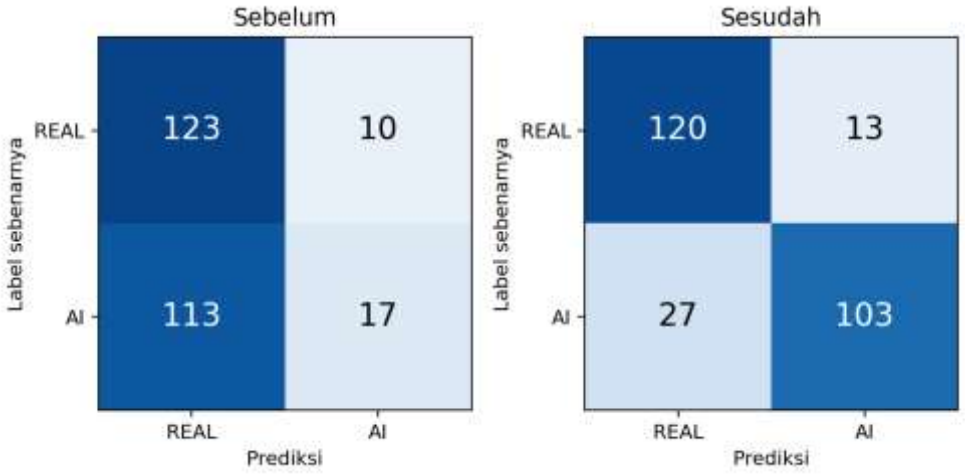
Fine-tuning memilih epoch ketujuh dengan skor validasi gabungan 0,7991. Kalibrasi fusion menghasilkan temperature 1,1065 dan threshold 0,70. Pada 263 citra uji domain baru yang dinormalisasi, akurasi fusion meningkat dari 53,23% menjadi 84,79%, sedangkan macro-F1 meningkat dari 0,4416 menjadi 0,8473. Spatial meningkat dari akurasi 55,13% menjadi 83,65%, dan spectral dari 30,80% menjadi 48,29%. Hasil ketiga head disajikan pada Tabel 4. Peningkatan terbesar secara absolut terjadi pada fusion, yaitu 31,56 poin persentase. Meskipun demikian, hasil akhir spatial dan fusion berdekatan, sehingga selisih akurasi 1,14 poin persentase antara keduanya perlu diperlakukan sebagai perbedaan deskriptif.

Tabel 4. Performa kepala klasifikasi pada domain baru yang dinormalisasi

Head	Akurasi awal	Akurasi akhir	Macro-F1 awal	Macro-F1 akhir
Spatial	55,13%	83,65%	0,4663	0,8362
Spectral	30,80%	48,29%	0,2659	0,4823
Fusion	53,23%	84,79%	0,4416	0,8473

Pada fusion, FPR berubah dari 0,075 menjadi 0,098, sedangkan FNR turun dari 0,869 menjadi 0,208. Penurunan FNR menunjukkan bahwa model lebih banyak mengenali citra AI yang sebelumnya terlewatkan, dengan kenaikan FPR yang relatif lebih kecil. Confusion matrix sebelum dan sesudah fine-tuning ditunjukkan pada Gambar 3. Sebelum fine-tuning, model mengklasifikasikan 17 citra AI secara benar dan melewatkan 113 citra AI. Sesudah fine-tuning, jumlah true positive meningkat menjadi 103 dan false negative turun menjadi 27, sementara false positive berubah dari 10 menjadi 13.

DOI: https://doi.org/10.69693/ijmst.v5i1.14703
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)



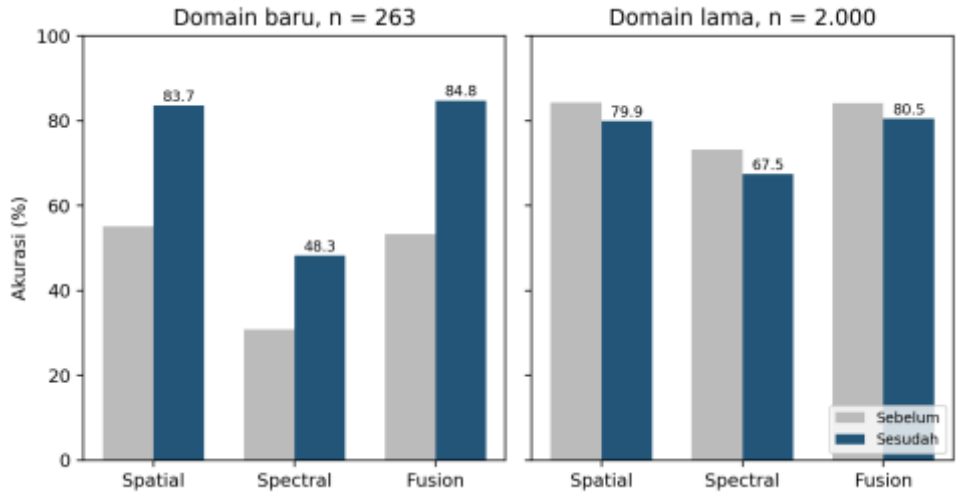
Gambar 3. Confusion matrix fusion sebelum dan sesudah fine-tuning

Perubahan prediksi pada citra yang sama diuji dengan McNemar exact. Sebanyak 92 citra berubah dari salah menjadi benar, sedangkan sembilan citra berubah dari benar menjadi salah; 31 citra tetap salah dan 131 tetap benar. Hasil uji menghasilkan $p < 0,001$. Nilai tersebut mendukung perubahan performa fusion sebelum dan sesudah fine-tuning pada sampel uji berpasangan. Uji ini tidak membuktikan bahwa fusion secara statistik lebih unggul daripada spatial, karena perbandingan antarkepala belum diuji menggunakan analisis prediksi berpasangan khusus.

Pada versi citra asli tanpa normalisasi file JPEG tambahan, fusion mencapai akurasi 85,17% dan macro-F1 0,8514. Nilai tersebut berbeda sekitar 0,38 poin persentase dari hasil versi yang dinormalisasi. Kedekatan kedua hasil menunjukkan bahwa performa akhir pada 263 citra uji relatif serupa pada dua versi prapemrosesan yang dievaluasi. Meskipun demikian, kedua versi berasal dari citra yang sama sehingga tidak boleh dianggap sebagai replikasi independen.

3.4 Regresi pada Domain Lama dan Trade-Off Adaptasi

Peningkatan performa pada domain baru disertai penurunan akurasi pada domain lama. Pada 2.000 citra external reference, akurasi fusion turun dari 84,10% sebelum fine-tuning menjadi 80,55% sesudah fine-tuning, atau berkurang 3,55 poin persentase. Macro-F1 fusion berubah dari 0,8377 menjadi 0,8004. Spatial turun dari 84,25% menjadi 79,90%, sedangkan spectral turun dari 73,15% menjadi 67,45%. Perubahan kedua domain dirangkum pada Gambar 4.



Gambar 4. Perubahan akurasi pada domain baru dan domain lama

DOI: https://doi.org/10.69693/ijmst.v5i1.14703
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Hasil ini memperlihatkan kompromi adaptasi domain: pembaruan parameter meningkatkan kesesuaian model terhadap distribusi data baru, tetapi sebagian kemampuan pada domain sebelumnya menurun. Penggunaan replay belum sepenuhnya mencegah regresi tersebut. Dengan demikian, peningkatan pada domain baru tidak dapat dipandang terpisah dari pemeliharaan performa domain lama. Pengembangan berikutnya dapat mengevaluasi komposisi replay, pembukaan parameter backbone, dan strategi pembelajaran bertahap menggunakan pengukuran pada kedua domain secara bersamaan.

3.5 Pengujian White Box dan Implikasi Temuan

Sebelas kasus pengujian white box dinyatakan lulus. Empat kasus pertama memeriksa jalur dasar fungsi fusi dan mencakup kondisi evaluasi, training tanpa dropout, training dengan cabang yang tetap tersedia, serta pemulihan cabang spatial ketika seluruh cabang terhapus. Kompleksitas siklomatik fungsi fusi adalah $V(G) = 4$, sehingga empat jalur dasar tersebut menjadi dasar pengujian alur kontrol. Kasus lain memeriksa ortogonalitas DCT, kestabilan transformasi pada citra konstan, dimensi embedding dan logit, pemisahan masukan antarkepala, aliran gradien, pemilihan threshold, dan keselarasan label dengan logit pada proses penggabungan kalibrasi. Ringkasan hasil ditampilkan pada Tabel 5.

Tabel 5. Ringkasan pengujian white box

Kelompok pengujian	Kasus	Hasil
Jalur dasar fungsi fusi	WB-01–04	4 jalur diperiksa; V(G) = 4; lulus
Transformasi dan kestabilan	WB-05–06	DCT ortogonal; tidak ada NaN/Inf; lulus
Dimensi dan pemisahan head	WB-07–08	Dimensi sesuai; masukan antarkepala terpisah; lulus
Aliran gradien	WB-09	Gradien finite dan tidak nol; lulus
Pencarian threshold	WB-10	Aturan skor terbaik dan seri sesuai; lulus
Penggabungan kalibrasi	WB-11	Label dan logit selaras; lulus

Secara empiris, integrasi fitur spasial dan forensik dapat dijalankan dalam satu arsitektur dan memberikan keuntungan numerik pada sebagian besar kondisi benchmark. Temuan ini sejalan dengan penelitian yang memanfaatkan frequency fingerprints dan kombinasi spasial-frekuensi untuk deteksi citra generatif (Alam et al., 2024; Wißmann et al., 2024; Bhattacharjee et al., 2026). Namun, hasil yang berbeda pada noise tinggi juga menegaskan perlunya menguji robustness berdasarkan jenis gangguan, bukan hanya melaporkan satu nilai akurasi agregat. Penelitian Nourji et al. (2025) dan Liu et al. (2026) juga menyoroti bahwa penggunaan petunjuk frekuensi tidak serta-merta menghilangkan persoalan generalisasi.

Dibandingkan penelitian berbasis CNN-Dilated, GLCM-CNN, atau deteksi inkonsistensi tekstur wajah, penelitian ini memperluas evaluasi pada citra sintetis dengan lima representasi forensik dan benchmark pascapemrosesan yang terstruktur (Taufiqurrahman et al., 2025; Hardiyanto & Akbar, 2025; Raharjo et al., 2025). Integrasi ConvNeXt juga mengikuti perkembangan ConvNet modern dan penelitian hybrid yang memanfaatkan backbone sejenis (Liu et al., 2022; Woo et al., 2023; Ilyas et al., 2024; Amarasinghe & Arachchi, 2025). Meskipun demikian, perbandingan hasil dengan penelitian lain tidak dapat digunakan untuk menyatakan model ini lebih unggul karena dataset, pembagian data, generator, dan protokol pengujian berbeda.

Keterbatasan utama penelitian adalah belum tersedianya ablation terkontrol untuk mengukur kontribusi individual FFT, DCT, wavelet Haar, fase FFT, dan SRM. Selain itu, keunggulan fusion terhadap spatial pada benchmark masih berupa perbandingan numerik, bukan bukti signifikansi statistik. Pengujian McNemar hanya menilai perubahan prediksi fusion sebelum dan sesudah fine-tuning. Pengujian white box juga membuktikan kesesuaian fungsi dan konfigurasi yang diuji, bukan jaminan bahwa model akan selalu benar pada seluruh data operasional. Keterbatasan tersebut perlu dipertimbangkan sebelum model diterapkan pada sistem verifikasi media berskala luas.

4. Kesimpulan

Penelitian ini berhasil mengintegrasikan ConvNeXt-Tiny dengan lima representasi forensik, yaitu magnitudo FFT, DCT, wavelet Haar, fase FFT, dan residual SRM, melalui gated residual fusion yang menghasilkan kepala

DOI: https://doi.org/10.69693/ijmst.v5i1.14703
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

klasifikasi spatial, spectral, dan fusion. Pada benchmark model dasar, fusion mencapai akurasi 93,00% pada citra bersih dan mengungguli spatial pada 18 dari 22 kondisi pengujian. Keuntungan terbesar terjadi pada resize 0,05 sebesar 1,69 poin persentase, sedangkan pada noise sigma 0,10 fusion lebih rendah 1,50 poin persentase. Fine-tuning meningkatkan akurasi fusion pada 263 citra domain baru yang dinormalisasi dari 53,23% menjadi 84,79% dan macro-F1 dari 0,4416 menjadi 0,8473; uji McNemar exact menghasilkan $p < 0,001$ untuk perubahan sebelum dan sesudah fine-tuning. Namun, akurasi pada 2.000 citra domain lama menurun dari 84,10% menjadi 80,55%. Seluruh sebelas kasus white box lulus, termasuk empat jalur dasar fungsi fusi dengan kompleksitas siklomatik empat. Hasil menunjukkan bahwa fusi spektral memberikan manfaat yang bergantung pada kondisi, sedangkan adaptasi domain menimbulkan kompromi antardomain. Penelitian lanjutan perlu melakukan ablation terkontrol, uji berpasangan antarkepala, memperluas variasi generator dan sumber citra, serta mengevaluasi biaya inferensi sebelum penerapan operasional.

Reference

1. Alam, I., Muneer, M. S., & Woo, S. S. (2024). UGAD: Universal generative AI detector utilizing frequency fingerprints. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 4332–4340). https://doi.org/10.1145/3627673.3680085
2. Amarasinghe, H. M. S., & Arachchi, S. P. K. (2025). Deepfake detection using a hybrid deep learning approach with Swin Transformers and ConvNeXt. In Proceedings of the International Research Conference on Smart Computing and Systems Engineering (SCSE 2025). https://doi.org/10.1109/SCSE65633.2025.11030955
3. Bergmann, S., Moussa, D., Brand, F., Kaup, A., & Riess, C. (2024). Forensic analysis of AI-compression traces in spatial and frequency domain. Pattern Recognition Letters, 180, 41–47. https://doi.org/10.1016/j.patrec.2024.02.015
4. Bhattacharjee, A., Islam, K., Anan, K., Intesher, A., Fuad, A. A., Saha, U., & Imtiaz, H. (2026). CAE-Net: Generalized deepfake image detection using convolution and attention mechanisms with spatial and frequency domain features. Journal of Visual Communication and Image Representation, 115, 104679. https://doi.org/10.1016/j.jvcir.2025.104679
5. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023). On the detection of synthetic images generated by diffusion models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1–5). https://doi.org/10.1109/ICASSP49357.2023.10095167
6. Hardiyanto, A., & Akbar, M. (2025). Klasifikasi citra sintetis hasil model difusi menggunakan Gray Level Co-Occurrence Matrix (GLCM) dan Convolutional Neural Network (CNN). JIKO (Jurnal Informatika dan Komputer), 9(3), 572–582. https://doi.org/10.26798/jiko.v9i3.2033
7. Ilyas, H., Javed, A., & Malik, K. M. (2024). ConvNext-PNet: An interpretable and explainable deep-learning model for deepfakes detection. In 2024 IEEE International Joint Conference on Biometrics (IJCB). https://doi.org/10.1109/IJCB62174.2024.10744484
8. Liu, C., Zhu, T., Zhou, W., & Zhao, W. (2026). Frequency bias matters: Diving into robust and generalized deep image forgery detection. IEEE Transactions on Dependable and Secure Computing, 23(2), 3636–3651. https://doi.org/10.1109/TDSC.2025.3639022
9. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11976–11986). https://doi.org/10.1109/CVPR52688.2022.01167
10. Nourji, S., Subramanian, T., Pakala, S., & Yang, E. (2025). Evaluating convolutional neural networks for synthetic image detection in the frequency domain. In 2025 IEEE International Conference on AI and Data Analytics (ICAD). https://doi.org/10.1109/ICAD65464.2025.11114037
11. Raharjo, T., Adristi, F. I., Romadhona, F. Y., Syahputra, R. R., Halim, M. Y., Rachman, M. A., Marjianto, R. N., Santicho, D., Kusuma, P., & Ramadhani, E. (2025). Analisis forensik deepfake berbasis Convolutional Neural Network (CNN) untuk deteksi inkonsistensi tekstur dan pola pada citra wajah. JATI (Jurnal Mahasiswa Teknik Informatika), 9(2), 2731–2738. https://doi.org/10.36040/jati.v9i2.13058
12. Ricker, J., Damm, S., Holz, T., & Fischer, A. (2024). Towards the detection of diffusion model deepfakes. In Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP) (Vol. 4, pp. 446–457). https://doi.org/10.5220/0012422000003660
13. Taufiqurrahman, R. F., Anggraeny, F. T., & Sihananto, A. N. (2025). Penerapan arsitektur CNN-Dilated untuk deteksi manipulasi citra deepfake. ILKOMNIKA: Journal of Computer Science and Applied Informatics, 7(3), 495–506. https://doi.org/10.28926/ilkomnika.v7i3.821
14. Wißmann, A., Zeiler, S., Nickel, R. M., & Kolossa, D. (2024). Whodunit: Detection and attribution of synthetic images by leveraging model-specific fingerprints. In Proceedings of the ACM Workshop on Multimedia AI against Disinformation (MAD '24) (pp. 65–72). https://doi.org/10.1145/3643491.3660280
15. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., & Xie, S. (2023). ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16133–16142). https://doi.org/10.1109/CVPR52729.2023.01548
16. Zhao, H., Wei, T., Zhou, W., Zhang, W., Chen, D., & Yu, N. (2021). Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2185–2194). https://doi.org/10.1109/CVPR46437.2021.00222

DOI: https://doi.org/10.69693/ijmst.v5i1.14703
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)