P-ISSN : 2549-7480
E-ISSN : 2549-6301

JURNAL ILMIAH INFORMATIKA
E-ISSN : 2549-6301
ISSN : 2549-7480
Faculty of Sains and Technology, Ibrahimy University
S4-Accredited : 105/E/KPT/2022
Journal page is available to : https://journal.ibrahimy.ac.id/index.php/JIMI

ANALISIS PERBANDINGAN VISION TRANSFORMER DAN CONVOLUTIONAL NEURAL NETWORK DALAM DETEKSI CITRA DEEPFAKE WAJAH MANUSIA DENGAN EVALUASI ROBUSTNESS DAN INTERPRETABILITAS MODEL

Nomensen Mangaratua Siregar[1)], Avip Kurniawan[2)], Nur Hikmah[3)].

[1] *Teknik Informatika, Universitas Krisnadwipayana*
[2] *Teknik Informatika, Universitas Krisnadwipayana*
[3] *Teknik Informatika, Universitas Krisnadwipayana*

email: [1] nomensen85@gmail.com, [2] avip.kurniawan@unkris.ac.id, [3] nurhikmah@unkris.ac.id

ARTICLE INFO

Article History:
Recieved : 9 Oktober 2026
Accepted : 2 November 2026
Published : 15 November 2026

Keywords:
Deepfake;
Vision Transformer;
CNN;
ResNet-50;
Robustness;

IEEE style in citing this article:
N.M. Siregar, A. Kurniawan, Nurhikmah, "Analisis Perbandingan Vision Transformer Dan Convolutional Neural Network Dalam Deteksi Citra Deepfake Wajah Manusia Dengan Evaluasi Robustness Dan Interpretabilitas", *jurnal.ilmiah.informatika*, vol. 11, no. 2, pp. 185-xx, Des. 2026.

Corresponding Author:
Nomensen Mangaratua Siregar
Teknik Informatika,
Universitas Krisnadwipayana

ABSTRACT

The spread of AI-generated deepfakes poses a growing threat to public information integrity, particularly face manipulations that are hard to distinguish from real images. However, existing detection evaluations often focus only on baseline accuracy, overlooking robustness against image degradation, model interpretability, and real-world applicability. This study compares Vision Transformer (ViT-Base) and Convolutional Neural Network (ResNet-50) using a StyleGAN3-based dataset across four dimensions: baseline accuracy, robustness against seven image degradation scenarios, interpretability via Attention Rollout and Grad-CAM, and real-world validation using a FastAPI interface with a soft-voting ensemble. Unlike previous studies that mostly focus on accuracy or hybrid architectures, this study compares the two pure architectures in a single, unified evaluation design. On clean data, ResNet-50 achieves higher accuracy (97.73%) than ViT-Base (94.67%). However, ViT-Base is far more robust to image degradation, with a maximum relative accuracy drop of 24.21% compared with 48.52% for ResNet-50. Grad-CAM and Attention Rollout correctly predicted 10 and 9 of 10 test samples, respectively, while the soft-voting ensemble corrected all ViT-Base errors on 10 out-of-distribution real-world samples. Overall, this study shows a clear performance trade-off between the two architectures, so model selection should match the intended operating environment.

© 2026 Jurnal Ilmiah Informatika (Scientific Informatics Journal) with CC BY NC licence

P-ISSN : 2549-7480
N. M. Siregar dkk/ JIMI 11 (2) pp. 185-199 E-ISSN : 2549-6301

1. PENDAHULUAN

Dalam beberapa tahun terakhir, teknologi *deepfake* telah berkembang dari isu akademis menjadi ancaman nyata bagi individu, korporasi, hingga proses demokrasi. Pada Januari 2024, kantor cabang Hong Kong sebuah perusahaan teknik multinasional mengalami kerugian sekitar 25,6 juta dolar AS setelah seorang karyawan keuangan ditipu melalui panggilan video yang menampilkan rekayasa wajah dan suara pimpinan perusahaan [1]. Di Indonesia, penyalahgunaan serupa muncul dalam bentuk video manipulasi tokoh publik yang membingungkan masyarakat menjelang Pemilu 2024 [2] serta penipuan bermodus bantuan pemerintah [3]. Ancaman ini juga menyasar citra statis, misalnya manipulasi foto mahasiswi berbasis AI di lingkungan kampus yang berdampak psikologis serius bagi korban [4]. Rangkaian kasus tersebut menunjukkan bahwa *deepfake* merugikan setidaknya tiga lapisan sekaligus: finansial, informasi publik, dan martabat individu.

Dari sisi teknis, sintesis wajah berbasis *Generative Adversarial Network* (GAN) telah mencapai realisme yang sangat tinggi dan mencakup empat kategori manipulasi, yaitu *face swap*, *face reenactment*, *face synthesis*, dan *face attribute manipulation* [5], [6], [7]. StyleGAN3 merupakan salah satu arsitektur generatif termutakhir dengan pendekatan *alias-free* yang menghasilkan citra hampir bebas artefak visual [8]. Survei terbaru menegaskan bahwa kemajuan metode generasi dan deteksi berlangsung sangat cepat, sehingga kajian deteksi harus terus diperbarui [9], [10]. Kondisi ini menuntut sistem deteksi otomatis yang tidak hanya akurat pada kondisi ideal, tetapi juga tangguh terhadap kualitas citra dunia nyata.

Convolutional Neural Network (CNN) selama ini menjadi pendekatan utama deteksi *deepfake* karena mampu mempelajari fitur lokal secara hierarkis [11], tetapi cenderung mengabaikan hubungan global antarbagian wajah dan terbukti rentan terhadap degradasi citra seperti kompresi JPEG, *Gaussian noise*, dan *motion blur* [12], [13]. *Vision Transformer* (ViT) yang mengadaptasi mekanisme *self-attention* ke domain citra [14] mampu menangkap *long-range dependencies* antar-*patch* sejak lapisan pertama dan menunjukkan ketahanan yang lebih baik terhadap gangguan citra [12], [15].

Sejumlah penelitian telah memanfaatkan kedua pendekatan tersebut untuk deteksi *deepfake*. Heo dkk. menunjukkan bahwa ViT dengan *distillation* mengurangi *overfitting* dan *false negative* dibandingkan pendekatan berbasis CNN pada dataset DFDC [16]. Coccomini dkk. membandingkan ViT dan CNN pada skenario lintas-manipulasi untuk menguji kemampuan generalisasi [17], dan kombinasi EfficientNet dengan ViT juga telah diusulkan untuk deteksi video [18]. Pendekatan *ensemble* ViT dan CNN dilaporkan lebih tangguh terhadap masukan berkualitas rendah serta lebih baik pada dataset yang berbeda dari data pelatihan [19]. Selain itu, Tex-ViT mengusulkan detektor berbasis tekstur dengan *dual-branch cross-attention* yang dirancang generalisabel dan tangguh [20], sementara Hasan dkk. melakukan perbandingan terkontrol CNN dan ViT yang menyimpulkan bahwa *self-attention* global meningkatkan ketangguhan, tetapi masih menyisakan kesenjangan generalisasi lintas domain [21]. Pada kondisi dunia nyata, *benchmark in-the-wild* menunjukkan penurunan performa signifikan pada detektor yang diuji dengan konten *deepfake* yang benar-benar beredar [22].

DOI : https://doi.org/10.35316/jimi.v11i2.185-199 © 2026 Jurnal Ilmiah Informatika with CC BY NC licence

Meskipun demikian, penelitian-penelitian tersebut umumnya berfokus pada akurasi pada data bersih, generalisasi lintas dataset, atau usulan arsitektur hibrida. Perbandingan langsung antara arsitektur murni CNN dan ViT yang sekaligus mencakup akurasi *baseline*, ketangguhan terhadap gangguan sintetis terkontrol, interpretabilitas keputusan model, dan validasi pada data di luar distribusi pelatihan melalui antarmuka yang dapat diakses pengguna masih terbatas. Sepengetahuan penulis, keempat aspek tersebut belum dikaji secara simultan dalam satu rancangan eksperimen yang konsisten.

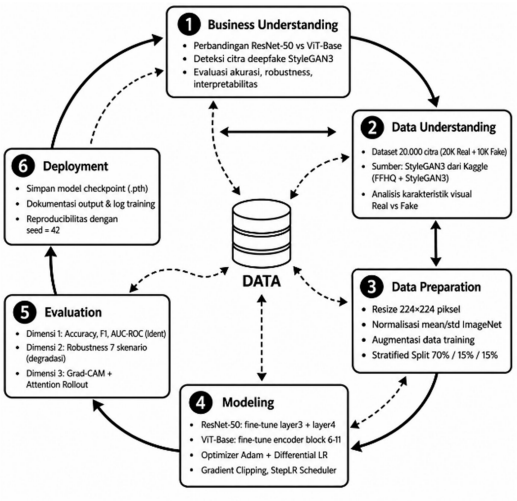
Oleh karena itu, penelitian ini bertujuan membandingkan ResNet-50 sebagai representasi CNN dan ViT-Base sebagai representasi *Vision Transformer* dalam mendeteksi citra wajah *deepfake* hasil StyleGAN3 melalui empat dimensi evaluasi: (1) akurasi pada data uji bersih, (2) ketangguhan terhadap tujuh skenario degradasi citra, (3) interpretabilitas keputusan model menggunakan Grad-CAM dan *Attention Rollout*, dan (4) validasi generalisasi pada citra *real-world out-of-distribution* melalui antarmuka web berbasis FastAPI dengan skema *ensemble soft voting*. Kontribusi penelitian ini adalah gambaran *trade-off* yang terukur antara kedua arsitektur sebagai dasar memilih model sesuai lingkungan penerapannya.

2. METODE PENELITIAN

Penelitian ini merupakan penelitian terapan dengan pendekatan kuantitatif-komparatif yang membandingkan ResNet-50 (representasi CNN) dan ViT-Base (representasi *Vision Transformer*) dalam mendeteksi citra *deepfake* wajah manusia pada empat dimensi evaluasi: akurasi pada data bersih, ketangguhan terhadap degradasi citra, interpretabilitas keputusan model, dan validasi generalisasi pada data di luar distribusi pelatihan (*out-of-distribution*).

2.1 Kerangka Kerja CRISP-DM

Penelitian ini mengikuti kerangka *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang terdiri dari enam fase, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*, sebagaimana Gambar 1. Fase *Business Understanding* mendefinisikan permasalahan deteksi *deepfake* beserta dampaknya, merumuskan tujuan penelitian secara terukur, dan menetapkan kriteria keberhasilan pada Tabel 4. Kelima fase lainnya diuraikan pada subbab berikut.



Gambar 1. Kerangka Kerja CRISP-DM

Dari Gambar 1 dapat dilihat bahwa fase-fase CRISP-DM berjalan iteratif, sehingga hasil evaluasi dapat menjadi masukan bagi perbaikan fase sebelumnya.

2.2 Dataset dan Pembagian Data

Dataset utama adalah Real vs Fake Faces (StyleGAN3) yang tersedia publik di Kaggle (troykueh/real-vs-fake-faces-stylegan3), terdiri dari 20.000 citra wajah beresolusi tinggi: 10.000 citra *Real* bersumber dari FFHQ (*Flickr-Faces-HQ*) dan 10.000 citra *Fake* hasil generasi

P-ISSN : 2549-7480
N. M. Siregar dkk/ JIMI 11 (2) pp. 185-199 E-ISSN : 2549-6301

StyleGAN3 [8] . Selain itu, dikumpulkan 10 sampel data primer untuk validasi *real-world*, yaitu 5 foto wajah otentik yang diambil dengan persetujuan subjek dan 5 foto hasil *face swap* berbasis InsightFace (melalui HuggingFace Space) dengan wajah target dari thispersondoesnotexist.com. Sumber generatif ini sengaja dibedakan dari StyleGAN3 agar benar-benar menguji generalisasi *out-of-distribution*. Dataset utama dibagi menggunakan *stratified split* dengan *random seed* 42 sebagaimana Tabel 1.

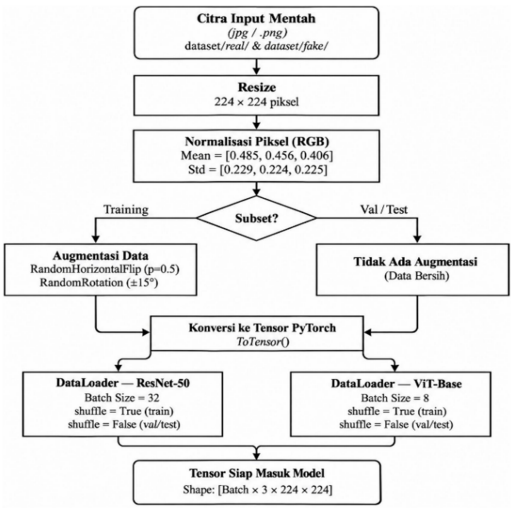
Tabel 1. Distribusi dataset Real vs Fake Faces

Subset	Real	Fake	Total
Training (70%)	7.000	7.000	14.000
Validasi (15%)	1.500	1.500	3.000
Pengujian (15%)	1.500	1.500	3.000
Total	10.000	10.000	20.000

Dari Tabel 1 dapat dilihat bahwa subset *training* digunakan untuk melatih bobot model, subset validasi untuk memantau performa per *epoch* dan menyimpan model terbaik, sedangkan subset pengujian dipisahkan sepenuhnya dari proses pelatihan dan validasi sehingga hasilnya merepresentasikan estimasi performa yang tidak bias.

2.3 Preprocessing Data

Seluruh citra melalui *pipeline* praproses seragam menggunakan *torchvision.transforms* sebagaimana Gambar 2: *resize* ke 224×224 piksel (interpolasi *bilinear*), konversi ke *tensor*, dan normalisasi dengan statistik ImageNet-1K (*mean* = [0,485; 0,456; 0,406], *std* = [0,229; 0,224; 0,225]). Augmentasi hanya diterapkan pada subset *training* berupa *RandomHorizontalFlip* (p = 0,5) dan *RandomRotation* (±15°), karena keduanya tidak mengubah keaslian wajah. *Batch size* ditetapkan 32 untuk ResNet-50 dan 8 untuk ViT-Base karena ViT-Base menyimpan *attention weights* dari 12 blok *encoder* sehingga membutuhkan VRAM lebih besar.



Gambar 2. Pipeline Preprocessing Data

Gambar 2 menunjukkan bahwa perlakuan praproses identik untuk kedua model, sehingga perbedaan hasil tidak berasal dari perbedaan masukan.

2.4 Arsitektur dan Pelatihan Model

Kedua model memakai strategi *partial fine-tuning* dari bobot *pretrained* ImageNet [23]: lapisan awal yang mempelajari fitur universal dibekukan, dan hanya lapisan akhir yang diadaptasi ke domain deteksi *deepfake*. Strategi ini dipilih karena *linear probing* kurang akurat pada domain *deepfake*, sedangkan *full fine-tuning* berisiko *catastrophic forgetting* pada data yang relatif kecil (14.000 sampel).

ResNet-50 [24] merepresentasikan CNN, dengan kontribusi utama berupa *skip connection* yang mengatasi *vanishing gradient* pada jaringan dalam, sebagaimana Persamaan (1).

$$y = F(x) + x \quad \ldots(1)$$

Dengan y adalah keluaran *residual block*, x adalah masukan (identitas), dan F(x) adalah pemetaan residual yang dipelajari lapisan konvolusi. Model dimuat dengan bobot IMAGENET1K_V1; *layer3* dan *layer4* dilatih ulang (±40% parameter), sedangkan *conv1*, *layer1*, dan *layer2* dibekukan. *Classification head* baru berupa Linear(2048→512)–ReLU–Dropout(0,4)–Linear(512→2).

ViT-Base [14] memperlakukan citra sebagai urutan *patch* 16×16 piksel dengan token klasifikasi [CLS], dan menghitung relevansi antar-token melalui *scaled dot-product attention* pada Persamaan (2).

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \ldots(2)$$

Dengan Q, K, V berturut-turut adalah matriks *Query*, *Key*, dan *Value*, serta d_k adalah dimensi *Key*. Model dimuat dari HuggingFace (google/vit-base-patch16-224) dengan *attn_implementation* = 'eager' agar *attention weights* tersedia untuk *Attention Rollout*. *Embeddings* dan *encoder block* 0–5 dibekukan, sedangkan *block* 6–11 dan *LayerNorm* akhir dilatih ulang (±50% parameter). *Classification head* baru berupa Linear(768→256)–GELU–Dropout(0,4)–Linear(256→2). Konfigurasi pelatihan kedua model ditunjukkan pada Tabel 2.

Tabel 2. Konfigurasi hyperparameter pelatihan

Hyperparameter	ResNet-50	ViT-Base
Optimizer	Adam	Adam
LR Backbone	1×10^{-4}	1×10^{-4}
LR Classification Head	5×10^{-4}	5×10^{-4}
Weight Decay	1×10^{-4}	1×10^{-4}
Batch Size	32	8
Epoch	20	20
LR Scheduler	StepLR (step = 12, γ = 0,2)	StepLR (step = 12, γ = 0,2)
Gradient Clipping	max_nor m = 1,0	max_nor m = 1,0

Dari Tabel 2 dapat dilihat bahwa seluruh *hyperparameter* identik kecuali *batch size*, sehingga perbedaan performa yang teramati merupakan konsekuensi perbedaan arsitektur, bukan konfigurasi pelatihan.

2.5 Skenario Pengujian Robustness

Evaluasi *robustness* mengukur sejauh mana model mempertahankan kinerjanya pada kondisi citra yang tidak ideal [24]. Gangguan diterapkan secara *on-the-fly* pada *test set* yang sama dengan evaluasi data bersih, terdiri dari tiga jenis gangguan pada dua level intensitas (Tabel 3) ditambah satu kondisi *clean baseline*, sehingga totalnya tujuh skenario.

Tabel 3. Parameter pengujian robustness

Jenis Gangguan	Parameter	Representasi
Gaussian Noise	σ = 0,05 / 0,15	Noise sensor kamera
Motion Blur	Kernel 5×5 / 12×12	Gerakan kamera tidak stabil
Kompresi JPEG	QF = 50 / 30	Unggahan ke platform digital

Dari Tabel 3 dapat dilihat bahwa setiap gangguan merepresentasikan degradasi yang berbeda di dunia nyata, dan pengujian pada dua level intensitas memungkinkan pengamatan kurva degradasi secara bertahap. Ketangguhan dinyatakan sebagai penurunan akurasi relatif terhadap *baseline* pada Persamaan (3).

$$\text{Drop} = \left(\frac{\text{Acc}_{\text{clean}} - \text{Acc}_{\text{degradasi}}}{\text{Acc}_{\text{clean}}}\right) \times 100\% \quad \ldots(3)$$

2.6 Metrik Evaluasi dan Kriteria Keberhasilan

Performa model dievaluasi dari *confusion matrix* (TP, TN, FP, FN) dengan

DOI : https://doi.org/10.35316/jimi.v11i2.185-199
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

P-ISSN : 2549-7480
N. M. Siregar dkk/ JIMI 11 (2) pp. 185-199 E-ISSN : 2549-6301

kelas *Fake* sebagai kelas positif, diturunkan menjadi *Accuracy*, *Precision*, *Recall*, *F1-Score* [25], serta AUC-ROC yang mengukur kemampuan diskriminatif model pada berbagai ambang batas keputusan. Rumusnya ditunjukkan pada Persamaan (4) sampai (7).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \ldots(4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \ldots(5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \ldots(6)$$

$$\text{F1} - \text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \ldots(7)$$

Keberhasilan penelitian dinilai berdasarkan empat kriteria pada Tabel 4.

Tabel 4. Kriteria keberhasilan penelitian

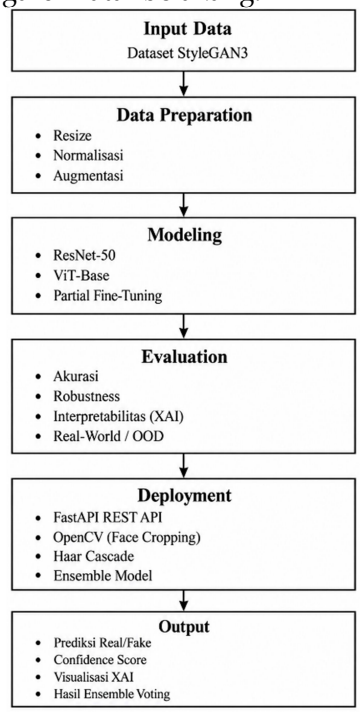
Kriteria	**Ambang Batas**
Kemampuan diskriminatif pada *clean test set*	AUC-ROC ≥ 0,85 (kedua model)
Ketangguhan terhadap degradasi citra	Penurunan akurasi relatif ≤ 30% dari *baseline* pada setiap skenario
Akurasi prediksi pada sampel uji interpretabilitas	≥ 70% sampel uji
Deteksi pada skenario *face swap*	Prediksi benar pada pengujian *real-world*

Dari Tabel 4 dapat dilihat bahwa kriteria pertama hingga ketiga dievaluasi pada dataset StyleGAN3, sedangkan kriteria keempat dievaluasi khusus pada data primer *real-world*.

2.7 Interpretabilitas dan *Deployment*

Interpretabilitas dievaluasi pada 10 sampel uji (5 *Real*, 5 *Fake*) menggunakan dua metode *post-hoc* yang bersifat *model-specific*. Grad-CAM [26] menghasilkan peta lokalisasi berbasis gradien pada lapisan konvolusi target ResNet-50 (*layer4[-1]*, resolusi 7×7), sedangkan *Attention Rollout* [27] mengakumulasikan bobot atensi dari seluruh *encoder block* ViT-Base secara rekursif untuk merekonstruksi aliran informasi menuju token [CLS].

Pada tahap *deployment*, kedua model diintegrasikan ke antarmuka web berbasis FastAPI sebagai layanan REST API dengan *endpoint* POST /predict/resnet50, POST /predict/vit, dan POST /predict/both (*ensemble soft voting*), sebagaimana Gambar 3. Setiap citra uji terlebih dahulu melewati fungsi *crop_face_with_margin* (OpenCV Haar Cascade Classifier, margin 25%) sebelum masuk ke *pipeline* praproses standar, sehingga model hanya mengenali artefak pada area wajah tanpa terpengaruh latar belakang.



Gambar 3. Alur Kerja Sistem

Gambar 3 menunjukkan alur dari unggahan citra, deteksi dan pemotongan wajah, praproses, hingga inferensi dan penyajian hasil klasifikasi beserta visualisasi XAI.

3. HASIL DAN PEMBAHASAN

Hasil disajikan menurut empat dimensi evaluasi, kemudian dinilai terhadap tujuan dan tolak ukur keberhasilan, dibandingkan dengan penelitian terdahulu, dan ditutup dengan pekerjaan yang belum berhasil beserta tindak lanjutnya.

3.1 Hasil Pelatihan dan Akurasi Baseline

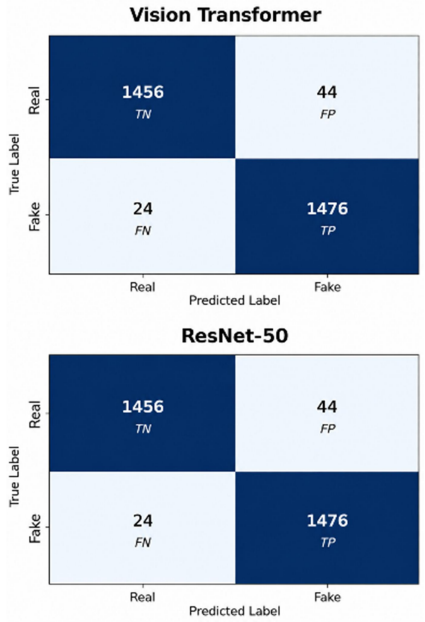
Kedua model dilatih 20 *epoch* dengan strategi *partial fine-tuning* yang identik. Keduanya mencapai *checkpoint* terbaik pada *epoch* ke-16 (Tabel 5). ResNet-50 belajar lebih stabil, dengan akurasi validasi 95,50% sudah pada *epoch* ke-4, sedangkan ViT-Base mengalami fluktuasi *validation loss* pada fase awal.

Tabel 5. Ringkasan hasil pelatihan pada epoch terbaik

Metrik	ResNet-50	ViT-Base
Epoch Terbaik	16	16
Train Accuracy	99,61%	98,60%
Validation Accuracy	98,40%	95,13%
Validation Loss	0,0710	0,2588
Generalization Gap	1,21 poin	3,47 poin

Dari Tabel 5 dapat dilihat bahwa *generalization gap* ResNet-50 (1,21 poin) lebih kecil daripada ViT-Base (3,47 poin), sehingga ResNet-50 tidak menunjukkan *overfitting* signifikan meskipun akurasi pelatihannya sangat tinggi.

Pada *clean test set* (3.000 citra), ResNet-50 menghasilkan TP = 1.476, TN = 1.456, FP = 44, FN = 24, sedangkan ViT-Base menghasilkan TP = 1.418, TN = 1.422, FP = 78, FN = 82 (Gambar 4). Metrik lengkap ditunjukkan pada Tabel 6.



Gambar 4. *Confusion Matrix* pada *Clean Test*

Tabel 6. Evaluasi performa pada data bersih

Metrik	ResNet-50	ViT-Base	Selisih (ResNet−ViT)
Accuracy	97,73%	94,67%	+3,06 poin
Precision	97,11%	94,79%	+2,32 poin
Recall	98,40%	94,53%	+3,87 poin
F1-Score	97,75%	94,66%	+3,09 poin
AUC-ROC	99,76%	98,94%	+0,82 poin

Dari Gambar 4 dan Tabel 6 dapat dilihat bahwa ResNet-50 unggul pada seluruh metrik, dengan selisih terbesar pada *Recall* (+3,87 poin) dan terkecil pada AUC-ROC (+0,82 poin). Kedua model memenuhi kriteria AUC-ROC ≥ 0,85 pada Tabel 4. Keunggulan ResNet-50 diduga berkaitan dengan kesesuaian antara konvolusi lokal dan artefak StyleGAN3 yang dapat berupa artefak frekuensi tinggi pada level piksel [8], tetapi dugaan

P-ISSN : 2549-7480
N. M. Siregar dkk/ JIMI 11 (2) pp. 185-199 E-ISSN : 2549-6301

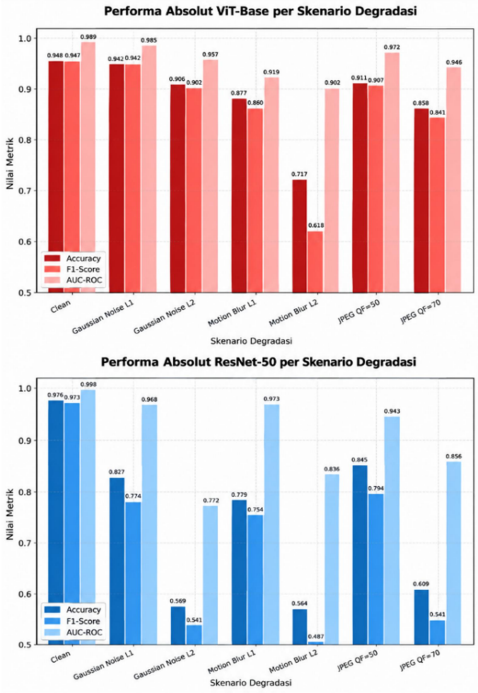
ini tidak diuji langsung dalam penelitian ini.

3.2 Ketangguhan terhadap Degradasi Citra

Hasil tujuh skenario degradasi ditunjukkan pada Tabel 7 dan Gambar 5. Penurunan akurasi dihitung relatif terhadap *baseline* sesuai Persamaan (3).

Tabel 7. Perbandingan penurunan akurasi pada tujuh skenario

Skenario	ResNet-50 Acc	ViT-Base Acc	Drop ResNet	Drop ViT
Clean (Baseline)	97,73%	94,67%	—	—
Gaussian $\sigma = 0,05$	82,73%	94,17%	15,35%	0,53%
Gaussian $\sigma = 0,15$	50,30%	90,63%	48,52%	4,27%
Motion Blur 5×5	76,93%	87,67%	21,30%	7,40%
Motion Blur 12×12	56,37%	71,73%	42,33%	24,21%
JPEG QF = 50	78,63%	90,87%	19,53%	4,02%
JPEG QF = 30	59,93%	85,83%	38,68%	9,33%

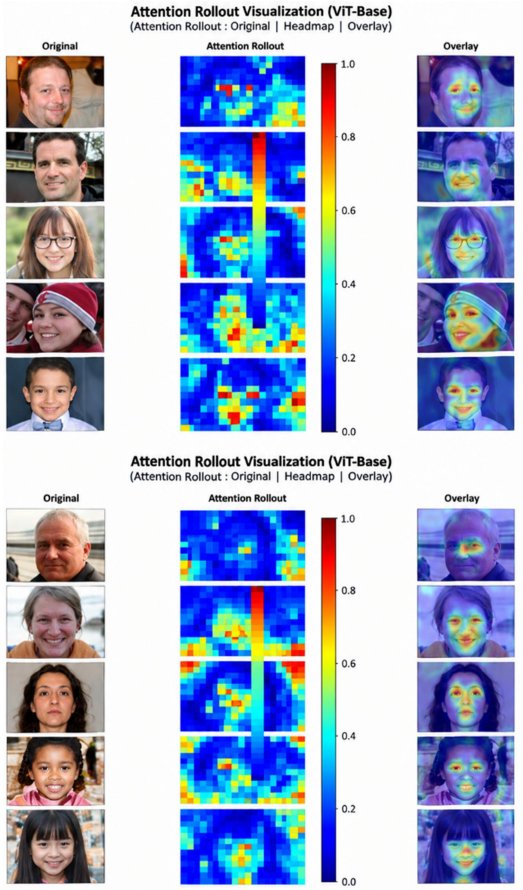


Gambar 5. Performa ResNet-50 dan ViT-Base pada pengujian *robustness*

Dari Tabel 7 dan Gambar 5 dapat dilihat bahwa ViT-Base lebih akurat daripada ResNet-50 pada keenam skenario gangguan, dan selisihnya membesar seiring intensitas gangguan. Pada *Gaussian noise*, selisih akurasi naik dari 11,44 poin ($\sigma = 0,05$) menjadi 40,33 poin ($\sigma = 0,15$). Rata-rata penurunan ResNet-50 adalah 30,95%, lebih dari dua kali rata-rata ViT-Base sebesar 8,29%. Pada $\sigma = 0,15$, akurasi ResNet-50 (50,30%) setara tebakan acak pada data seimbang dan F1-Score-nya jatuh hingga 0,0130, yang menandakan *collapse* prediksi ke satu kelas. ViT-Base menunjukkan *graceful degradation*, yaitu penurunan bertahap yang sebanding dengan intensitas gangguan, meskipun *motion blur* berat tetap menurunkan akurasinya cukup besar.

3.3 Intepretabilitas Model

Interpretabilitas dievaluasi pada 10 sampel uji (5 *Real*, 5 *Fake*). Grad-CAM pada ResNet-50 memprediksi benar 10 dari 10 sampel, sedangkan *Attention Rollout* pada ViT-Base memprediksi benar 9 dari 10 sampel, dengan satu sampel *Fake* diprediksi *Real* pada *confidence* 100%.



Gambar 6. Visualisasi interpretabilitas model pada sampel uji

Dari Gambar 6 dapat dilihat bahwa *heatmap* Grad-CAM terkonsentrasi pada area mata, alis, dan tepi wajah, sedangkan *Attention Rollout* menghasilkan distribusi atensi yang lebih global dan menyebar. Perbedaan ini sesuai dengan cara pandang lokal CNN dan global *Transformer*. Pengamatan ini kualitatif dan terbatas pada 10 sampel, sehingga belum membuktikan bahwa area yang disorot adalah dasar keputusan model.

3.4 Validasi Real World

Validasi *out-of-distribution* dilakukan pada 10 sampel data primer (5 *Real*, 5 *face swap* InsightFace) yang sumber generatifnya berbeda dari StyleGAN3 (Tabel 8).

Tabel 8. Perbandingan metrik evaluasi real-world

Metrik	ViT-Base	ResNet-50	*Ensemble* (*Soft Voting*)
Accuracy	80%	100%	100%
Precision (Fake)	80%	100%	100%
Recall (Fake)	80%	100%	100%
F1-Score (Fake)	80%	100%	100%

Dari Tabel 8 dapat dilihat bahwa ResNet-50 dan *ensemble soft voting* mencatat seluruh metrik 100%, sedangkan ViT-Base mencatat 80% dengan dua kesalahan klasifikasi. Salah satunya berkeyakinan tinggi, yaitu citra *Fake* diprediksi *Real* dengan *confidence* 99,9%. Kegagalan seperti ini berisiko pada penggunaan forensik karena keyakinan tinggi cenderung menurunkan kewaspadaan pengguna. *Ensemble soft voting* mengoreksi kedua kesalahan tersebut. Hasil ini merupakan studi kasus kualitatif dengan n = 10 dan bukan estimasi performa yang signifikan secara statistik.

3.5 Pencapaian Tujuan Penelitian

Tabel 9 merangkum pencapaian tiap tujuan terhadap tolak ukur pada Tabel 4.

Tabel 9. Pencapaian tujuan terhadap tolak ukur keberhasilan

Tujuan	Tolak ukur	Hasil	Status
Akurasi data bersih	AUC-ROC ≥ 0,85	ResNet-50 0,9976; ViT-Base 0,9894	Tercapai (keduanya)
Ketangguhan	Penurunan relatif ≤ 30% pada setiap skenario	ViT-Base maks. 24,21%; ResNet-50 maks. 48,52% (3 dari 6 skenario	Tercapai oleh ViT-Base; tidak oleh ResNet-50

P-ISSN : 2549-7480
N. M. Siregar dkk/ JIMI 11 (2) pp. 185-199 E-ISSN : 2549-6301

		melampaui)	
Interpretabilitas	Akurasi prediksi sampel XAI ≥ 70%	Grad-CAM 10/10; *Attention Rollout* 9/10	Tercapai (keduanya), terbatas pada penilaian kualitatif
Validasi *real-world*	Prediksi benar pada pengujian *real-world*	ResNet-50 dan *ensemble* 10/10; ViT-Base 8/10	Tercapai oleh ResNet-50 dan *ensemble*; tidak oleh ViT-Base

Dari Tabel 9 dapat dilihat bahwa tujuan utama penelitian, yaitu membandingkan kedua arsitektur pada empat dimensi, berhasil dilaksanakan dan menunjukkan *trade-off* yang jelas. Tidak ada model yang memenuhi semua tolak ukur: ResNet-50 gagal pada ketangguhan, ViT-Base gagal pada validasi *real-world*.

Penyebab kegagalan ResNet-50 pada dimensi ketangguhan diduga karena augmentasi pelatihan hanya berupa *flip* dan rotasi, tanpa *noise*, *blur*, atau kompresi, sehingga model tidak pernah melihat degradasi tersebut. Ini sejalan dengan pola bahwa penurunan membesar seiring intensitas gangguan. Kegagalan ViT-Base pada validasi *real-world* diduga berkaitan dengan *partial fine-tuning* pada 14.000 sampel yang belum cukup menstabilkan bobot atensi, sebagaimana tampak dari fluktuasi *validation loss* awal dan *generalization gap* yang lebih besar. Kedua penyebab ini belum diuji dengan eksperimen ablasi. Perlu dicatat pula bahwa tolak ukur interpretabilitas mengukur akurasi prediksi pada sampel XAI, bukan ketepatan area yang disorot terhadap area manipulasi sebenarnya.

3.6 Perbandingan dengan Penelitian Terdahulu

Tabel 10 membandingkan penelitian ini dengan studi terdahulu yang menggunakan ViT dan CNN.

Tabel 10. Perbandingan dengan penelitian terdahulu

Penelitian	**Metode dan data**	**Persamaan**	**Perbedaan dengan penelitian ini**
[16]	ViT dengan *distillation* dan fitur CNN; video DFDC	Sama-sama menilai ViT untuk deteksi *deepfake*	Penelitian ini memakai ViT-Base standar tanpa *distillation*, pada citra StyleGAN3
[17]	ViT dan EfficientNetV2; citra ForgeryNet; generalisasi lintas manipulasi	Membandingkan ViT dan CNN pada citra	Penelitian ini menambah degradasi terkontrol, XAI, dan validasi data primer
[19]	*Ensemble* ViT dan CNN dengan pemungutan suara berbobot; diuji pada dataset berbeda dari data latih	Sama-sama memakai *ensemble* ViT dan CNN	Penelitian ini memakai *soft voting* (rata-rata probabilitas), dan *ensemble* hanya satu dari empat dimensi
[21]	ResNet-18,	Perbandingan	Penelitian ini

DOI : https://doi.org/10.35316/jimi.v11i2.185-199 © 2026 Jurnal Ilmiah Informatika with CC BY NC licence

	EfficientNet-B0, ViT-B/16; Celeb-DF v2 dan FaceForensics++ (video)	terkontrol CNN dan ViT dengan konfigurasi seragam	berbasis citra tunggal dengan ResNet-50, ViT-Base, dan StyleGAN3, ditambah XAI dan data *real-world*

Dari Tabel 10 dapat dilihat bahwa persamaan utamanya adalah perbandingan CNN dan ViT, sedangkan perbedaannya ada pada jenis data (citra StyleGAN3 versus video atau dataset lain), cakupan dimensi evaluasi, dan arsitektur yang dibandingkan.

Dari sisi hasil, temuan bahwa ViT lebih tangguh daripada CNN searah dengan [12] dan [21], yang melaporkan bahwa *self-attention* global meningkatkan ketangguhan, serta dengan [19], yang menyebut CNN rentan terhadap masukan berkualitas rendah. Keberhasilan *ensemble* pada data di luar distribusi juga searah dengan [19]. Perbedaannya, ResNet-50 di sini lebih akurat pada data bersih, sedangkan [16] melaporkan varian ViT mengungguli pendekatan berbasis CNN pada DFDC. Perbedaan ini diduga berasal dari jenis data (citra StyleGAN3 versus video), strategi *partial fine-tuning* dengan data pelatihan relatif kecil, serta rancangan model pada [16] yang berbeda dari ViT-Base standar. Hasil ResNet-50 yang sempurna pada data *real-world* tidak dapat dibaca sebagai bukti generalisasi umum, mengingat *benchmark in-the-wild* melaporkan penurunan performa detektor pada konten yang benar-benar beredar [22].

3.7 Pekerjaan yang Belum Berhasil dan Tindak Lanjut

Tabel 11 merangkum pekerjaan yang belum berhasil dan kemungkinan tindak lanjutnya.

Tabel 11. Pekerjaan belum berhasil, penyebab, dan tindak lanjut

Belum berhasil	Penyebab	Tindak lanjut
ResNet-50 tidak tangguh terhadap degradasi	Augmentasi pelatihan tanpa *noise*, *blur*, dan kompresi	*Robustness-aware training* dengan augmentasi JPEG, *noise*, dan *blur*
ViT-Base kurang akurat pada data bersih dan memiliki kegagalan berkeyakinan tinggi	Diduga data pelatihan terbatas pada *partial fine-tuning*	*Full fine-tuning* dengan data lebih besar dan kalibrasi *confidence*
Generalisasi baru diuji pada 10 sampel dan satu metode *face swap*	Keterbatasan data primer	Evaluasi pada FaceForensics++, Celeb-DF, dan generator lain, dengan sampel lebih besar
Interpretabilitas hanya kualitatif	Belum ada anotasi area manipulasi	Metrik lokalisasi kuantitatif atau penilaian ahli
Hasil dari satu pembagian data dan deteksi wajah Haar Cascade	Keterbatasan eksperimen	Pengulangan dengan beberapa *seed* dan detektor wajah seperti MTCNN atau RetinaFace

DOI : https://doi.org/10.35316/jimi.v11i2.185-199
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

P-ISSN : 2549-7480
N. M. Siregar dkk/ JIMI 11 (2) pp. 185-199 E-ISSN : 2549-6301

4. KESIMPULAN

Pada data bersih, ResNet-50 unggul konsisten di kelima metrik evaluasi (*Accuracy* 97,73%, *Precision* 97,11%, *Recall* 98,40%, *F1-Score* 97,75%, AUC-ROC 99,76%) dibandingkan ViT-Base (*Accuracy* 94,67%, *Precision* 94,79%, *Recall* 94,53%, *F1-Score* 94,66%, AUC-ROC 98,94%). Namun, pengujian ketangguhan terhadap tujuh skenario degradasi citra justru membalik hasil tersebut. ViT-Base konsisten lebih tangguh dengan penurunan akurasi relatif maksimum 24,21%, sedangkan ResNet-50 turun hingga 48,52% disertai *collapse* prediksi pada skenario *Gaussian noise* terberat. Pada evaluasi interpretabilitas, Grad-CAM pada ResNet-50 memprediksi benar 10 dari 10 sampel dengan *heatmap* yang terlokalisir pada area wajah, sedangkan *Attention Rollout* pada ViT-Base memprediksi benar 9 dari 10 sampel dengan distribusi atensi global yang melengkapi cara pandang lokal Grad-CAM. Pada 10 sampel *real-world out-of-distribution*, ResNet-50 mempertahankan akurasi 100%, sementara ViT-Base mencapai 80% dengan satu kegagalan berkeyakinan tinggi (*confidence* 99,9% pada prediksi keliru). *Ensemble soft voting* mampu mengoreksi seluruh kesalahan ViT-Base sehingga akurasinya mencapai 100%. Secara keseluruhan, temuan ini mengonfirmasi adanya *trade-off* yang jelas dan terukur antara CNN dan *Transformer* untuk deteksi *deepfake*: ResNet-50 lebih unggul pada data bersih dan interpretabilitas yang presisi untuk keperluan audit forensik, sedangkan ViT-Base jauh lebih andal terhadap degradasi citra yang lazim ditemui di dunia nyata, dan pendekatan *ensemble* mampu menggabungkan keunggulan keduanya sekaligus meredam kelemahan masing-masing.

Penelitian ini memiliki beberapa keterbatasan. Pelatihan hanya memakai satu generator (StyleGAN3), validasi *real-world* hanya mencakup 10 sampel dari satu metode *face swap*, dan hasil berasal dari satu pembagian data, sehingga temuan belum bisa digeneralisasi ke generator, dataset, atau kondisi lain. Secara praktis, hasil ini menunjukkan bahwa memilih detektor tidak cukup hanya dari akurasi pada data bersih. Untuk citra yang sering terkompresi dan buram, seperti yang diunggah ke media sosial, ViT-Base lebih aman, dan *ensemble* dapat dipakai untuk mengurangi kesalahan berkeyakinan tinggi.

Untuk penelitian selanjutnya, disarankan eksplorasi arsitektur hibrida CNN-*Transformer* seperti EfficientViT atau ConvNeXt-*Transformer* untuk mengkaji potensi performa yang unggul pada akurasi dan *robustness* sekaligus. Dataset pelatihan juga dapat diperluas dengan metode generasi *deepfake* lain seperti InsightFace, DALL-E, Midjourney, dan *First Order Motion Model* untuk meningkatkan generalisasi lintas generator, dan *robustness-aware training* melalui augmentasi JPEG, *Gaussian noise*, dan *blur* dapat diterapkan untuk menutup kesenjangan *robustness* ResNet-50 tanpa mengorbankan akurasi. Selain itu, *knowledge distillation* dari ViT-Base ke arsitektur yang lebih ringan dapat dicoba untuk *deployment* pada perangkat *mobile/edge computing*, dan deteksi wajah pada antarmuka web dapat ditingkatkan dengan mengganti Haar Cascade menggunakan MTCNN, RetinaFace, atau MediaPipe Face Detector. Sistem juga dapat diperluas ke deteksi berbasis video secara *real-time* melalui model seperti Meso4 atau 3D-CNN, serta dievaluasi lintas dataset yang lebih besar seperti FaceForensics++ atau Celeb-DF untuk mengukur generalisasi di luar konteks StyleGAN3.

DOI : https://doi.org/10.35316/jimi.v11i2.185-199
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

5. REFERENSI

[1] Heather Chen and Kathleen Magramo, "Finance worker pays out $25 million after video call with deepfake 'chief financial officer,'" CNN. Accessed: Nov. 03, 2026. [Online]. Available: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

[2] CNN Indonesia, "Kominfo Bongkar Cara Produksi Video Deepfake Jokowi Berbahasa Mandarin," CNN Indonesia. Accessed: Sep. 10, 2026. [Online]. Available: https://www.cnnindonesia.com/teknologi/20231026182710-192-1016431/kominfo-bongkar-cara-produksi-video-deepfake-jokowi-berbahasa-mandarin

[3] Kompas.com, "Kasus Video Deepfake Prabowo, Semua Bisa Jadi Korban AI." Accessed: Sep. 10, 2026. [Online]. Available: https://nasional.kompas.com/read/2025/01/24/07562761/kasus-video-deepfake-prabowo-semua-bisa-jadi-korban-ai

[4] ANTARA News, "Universitas Udayana investigasi kasus dugaan pelecehan seksual daring." Accessed: Sep. 10, 2026. [Online]. Available: https://www.antaranews.com/berita/4802537/universitas-udayana-investigasi-kasus-dugaan-pelecehan-seksual-daring

[5] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A Survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020, doi: https://doi.org/10.1016/j.inffus.2020.06.014.

[6] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes," Jul. 31, 2021, *Association for Computing Machinery*. doi: 10.1145/3425780.

[7] Z. Akhtar, "Deepfakes Generation and Detection: A Short Survey," *J. Imaging*, vol. 9, no. 1, p. 18, Jan. 2023, doi: 10.3390/jimaging9010018.

[8] T. Karras *et al.*, "Alias-Free Generative Adversarial Networks," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 852–863. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/076ccd93ad68be51f23707988e934906-Paper.pdf

[9] G. Pei *et al.*, "Deepfake Generation and Detection: A Benchmark and Survey," Mar. 2026.

[10] P. Liu, Q. Tao, and J. T. Zhou, "Evolving from Single-modal to Multi-modal Facial Deepfake Detection: Progress and Challenges," Apr. 2025.

[11] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00444-8.

[12] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, "Intriguing Properties of Vision Transformers," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 23296–23308. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/c404a5adbf90e09631678b13b05d9d7a-Paper.pdf

DOI : https://doi.org/10.35316/jimi.v11i2.185-199 © 2026 Jurnal Ilmiah Informatika with CC BY NC licence

P-ISSN : 2549-7480
E-ISSN : 2549-6301

[13] K. Shiohara and T. Yamasaki, "Detecting Deepfakes with Self-Blended Images," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18699–18708. doi: 10.1109/CVPR52688.2022.01816.

[14] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2010.11929

[15] K. Han *et al.*, "A Survey on Vision Transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, 2023, doi: 10.1109/TPAMI.2022.3152247.

[16] Y.-J. Heo, Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, "Deepfake Detection Scheme Based on Vision Transformer and Distillation," Apr. 2021.

[17] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, and G. Amato, "Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection," Jun. 2022, doi: 10.1145/3512732.3533582.

[18] N. and G. C. and F. F. Coccomini Davide Alessandro and Messina, "Combining EfficientNet and Vision Transformers for Video Deepfake Detection," in *Image Analysis and Processing – ICIAP 2022*, C. and L. M. and F. G. M. and T. F. Sclaroff Stan and Distante, Ed., Cham: Springer International Publishing, 2022, pp. 219–229.

[19] H. Ha, M. Kim, S. Han, and S. Lee, "Robust DeepFake Detection Method based on Ensemble of ViT and CNN," in *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, New York, NY, USA: ACM, Mar. 2023, pp. 1092–1095. doi: 10.1145/3555776.3577769.

[20] D. Dagar and D. Kumar Vishwakarma, "Tex-ViT: A Generalizable, Robust, Texture-based Dual-Branch Cross-Attention Deepfake Detector."

[21] S. A. Hasan, M. P. Hera, A. L. Mim, M. M. Jihad Hossain, H. Ali, and M. F. Bulbul, "Beyond Frame-Level Accuracy: Analyzing Video-Level Stability in Transformer-Based Deepfake Forensics," *International Journal of Intelligent Systems*, vol. 2026, no. 1, Jan. 2026, doi: 10.1155/int/6801006.

[22] N. A. Chandra *et al.*, "Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024," May 2026, [Online]. Available: http://arxiv.org/abs/2503.02857

[23] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," Nov. 2014, [Online]. Available: http://arxiv.org/abs/1411.1792

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[25] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

DOI : https://doi.org/10.35316/jimi.v11i2.185-199
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

P-ISSN: 2549-7480
N. M. Siregar dkk/ JIMI 11 (2) pp. 185-199 E-ISSN: 2549-6301

[27] S. Abnar and W. Zuidema, "Quantifying Attention Flow in Transformers."

DOI : https://doi.org/10.35316/jimi.v11i2.185-199
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence