JURNAL JTIK
(JURNAL TEKNOLOGI INFORMASI DAN KOMUNIKASI)
KITA Komunitas Informasi Teknologi Aceh
OPEN ACCESS
E-ISSN: 2580-1643
Volume 11, Nomor 1, Januari 2027 | Pages: 292-304 | DOI: https://doi.org/10.35870/jtik.v11i1.8767

Analisis Perbandingan Kinerja Arsitektur CNN dan Transformer dalam Klasifikasi Gambar *AI-Generated* dan Gambar Asli

David Yus Humisar Silalahi [1*], Junaidi [2], Eka Ari Endrawati [3]

[1*,2,3] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Krisnadwipayana, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta, Indonesia.

Corresponding Email: [1*]davidyushumisar@gmail.com.

Article History:
Received: 19 September 2026.
Revised: 8 October 2026.
Accepted: 14 October 2026.
Available Online: 1 November 2026.
Published: 1 January 2027.

Abstract

The advancement of generative artificial intelligence has enabled increasingly realistic facial images, creating a need for reliable methods to distinguish AI-generated images from real photographs. This study compares four deep-learning models for binary classification: ResNet50 and MobileNetV2 (CNNs), and Vision Transformer (ViT) and a Transformer implementation referred to as DeiT. The dataset was sourced from the 140k Real and Fake Faces collection on Kaggle; 28,000 images were reportedly used in a balanced split of 20,000 training, 4,000 validation, and 4,000 test images. Models were evaluated using accuracy, precision, recall, F1-score, and confusion matrices. DeiT achieved the highest reported test accuracy (86.98%), followed by MobileNetV2 (77.93%), ViT (71.83%), and ResNet50 (66.83%). Using the evaluasi-requested unrounded group aggregates, mean accuracy was 72.38% for CNNs and 79.40% for Transformers, a difference of 7.03 percentage points. These results apply only to the reported dataset and training configurations; differences in pretraining and model implementation mean the comparison should not be interpreted as an isolated causal effect of architecture. DeiT was deployed as a Gradio prototype.

Keywords: AI-generated image; CNN; DeiT; Gradio; Transformer.

Abstrak

Perkembangan kecerdasan buatan generatif memungkinkan pembuatan citra wajah yang semakin realistis sehingga diperlukan metode yang andal untuk membedakan citra buatan AI dan foto asli. Penelitian ini membandingkan empat model deep learning untuk klasifikasi biner, yaitu ResNet50 dan MobileNetV2 (CNN), serta Vision Transformer (ViT) dan implementasi Transformer yang disebut DeiT. Dataset bersumber dari 140k Real and Fake Faces di Kaggle; sebanyak 28.000 citra dilaporkan digunakan dengan pembagian seimbang: 20.000 citra pelatihan, 4.000 validasi, dan 4.000 pengujian. Evaluasi menggunakan accuracy, precision, recall, F1-score, dan confusion matrix. DeiT memperoleh accuracy pengujian tertinggi (86,98%), diikuti MobileNetV2 (77,93%), ViT (71,83%), dan ResNet50 (66,83%). Berdasarkan agregat kelompok yang diminta evaluasi, rata-rata accuracy CNN adalah 72,38%, sedangkan Transformer 79,40%, dengan selisih 7,03 poin persentase. Hasil ini terbatas pada dataset dan konfigurasi pelatihan yang dilaporkan; perbedaan prapelatihan dan implementasi model membuat hasil tidak dapat ditafsirkan sebagai efek arsitektur semata. DeiT diimplementasikan sebagai prototipe Gradio.

Kata Kunci: AI-generated image; CNN; Transformer; klasifikasi citra; Gradio.

This work is licensed under a Creative Commons Attribution 4.0 International License.

Copyright © 2027 by the authors. Published by Lembaga Komunitas Informasi Teknologi Aceh (KITA). All rights reserved.

E-ISSN: 2580-1643

1. Pendahuluan

Perkembangan kecerdasan buatan dalam beberapa tahun terakhir mendorong perubahan besar pada produksi dan pengolahan konten visual. Model generatif mampu menghasilkan gambar dengan kualitas yang semakin realistis, termasuk citra wajah yang secara kasat mata dapat menyerupai hasil fotografi. Kemampuan tersebut bermanfaat bagi industri kreatif, pendidikan, hiburan, pemasaran digital, dan pembuatan konten multimedia, tetapi pada saat yang sama menimbulkan persoalan autentisitas. Gambar yang dibuat oleh sistem AI dapat digunakan untuk menghasilkan informasi visual yang menyesatkan, memanipulasi identitas, atau membentuk bukti visual yang tidak sesuai dengan kondisi nyata. Akibatnya, kebutuhan untuk memverifikasi keaslian citra digital menjadi semakin penting (Gong & Li, 2024; Malik *et al*., 2022; Rana *et al*., 2022). Kesulitan utama pada proses verifikasi muncul karena peningkatan kualitas *generative AI* memperkecil perbedaan visual antara *AI-generated image* dan *real image*. Pada gambar asli, variasi tekstur, pencahayaan, *noise* sensor, bayangan, dan artefak kompresi terbentuk melalui proses akuisisi citra menggunakan kamera atau sensor optik. Sebaliknya, gambar *AI-generated* dihasilkan melalui model generatif yang menyintesis karakteristik visual berdasarkan pola pada data pelatihan. Walaupun beberapa citra sintetis masih menyisakan artefak pada rambut, tekstur kulit, transisi warna, latar belakang, atau domain frekuensi, perbedaan tersebut tidak selalu mudah dikenali secara manual.

Penelitian terkini menunjukkan bahwa detektor juga perlu menghadapi perubahan karakteristik generator dan pascaproses citra, sehingga generalisasi menjadi persoalan penting dalam identifikasi gambar sintetis (Bammey, 2023; Manisha *et al*., 2025). *Deep learning* menjadi pendekatan yang relevan karena mampu mempelajari representasi visual langsung dari data tanpa bergantung pada *feature engineering* manual. *Convolutional Neural Network* (CNN) merupakan salah satu arsitektur yang paling luas digunakan untuk klasifikasi citra karena operasi konvolusinya efektif dalam mengekstraksi fitur lokal seperti tepi, tekstur, bentuk, dan pola spasial. Penerapan *transfer learning* dan *data augmentation* dilaporkan dapat memperkuat *robustness* klasifikasi citra *AI-generated*, termasuk ketika data mengalami gangguan visual (Ranupadma *et al*., 2026). Implementasi CNN juga telah menunjukkan kemampuan yang baik dalam deteksi *deepfake* berbasis citra (Gulo *et al*., 2025). Pada keluarga ResNet, ekstraksi fitur mendalam maupun kombinasi dengan teknik klasifikasi lain mampu memberikan performa tinggi dalam membedakan citra autentik dan sintetis (Kusnaeni *et al*., 2025; Maulana *et al*., 2025; Rizqullah *et al*., 2026). Meskipun CNN kuat dalam ekstraksi fitur lokal, pendekatan ini memiliki keterbatasan ketika informasi diskriminatif tersebar pada bagian citra yang saling berjauhan. Perkembangan *Transformer* pada bidang *computer vision* menawarkan mekanisme yang berbeda melalui *self-attention*. *Vision Transformer* (ViT) membagi citra menjadi *patch* dan mempelajari hubungan antar-*patch* secara global sehingga model dapat menangkap keterkaitan antarelemen citra yang tidak bergantung pada *receptive field* lokal. Aria *et al.* (2025) menunjukkan bahwa ViT dapat digunakan untuk membedakan citra asli dan citra generatif, sedangkan Arshed *et al.* (2023) menerapkan ViT untuk deteksi *deepfake*. Kemampuan ViT dalam klasifikasi citra juga ditunjukkan dalam penelitian Tahyudin *et al.* (2024) dan Sihananto *et al.* (2024). Pengembangan lebih lanjut dari pendekatan tersebut menghasilkan *Data-efficient Image Transformer* (DeiT), yang bertujuan mempertahankan kekuatan *Transformer* dengan kebutuhan data pelatihan yang lebih efisien. Studi Erwin *et al.* (2025) menunjukkan bahwa DeiT dapat digunakan untuk membedakan gambar buatan AI dan manusia dengan performa yang kompetitif. Pada cakupan yang lebih luas, pendekatan berbasis *patch* dan *Transformer* juga dilaporkan efektif untuk klasifikasi *deepfake* multikelas, yang menunjukkan pentingnya pemodelan hubungan global pada citra sintetis (Arshed *et al*., 2024).

Temuan-temuan tersebut menunjukkan bahwa CNN maupun *Transformer* mampu menyelesaikan tugas klasifikasi citra. Namun, mekanisme pembelajaran fitur pada kedua arsitektur berbeda dan berpotensi menghasilkan performa yang berbeda pada *dataset* yang sama. Studi Aria *et al.* (2025) melaporkan *accuracy* ViT sebesar 98% dan EfficientNet-B0 sebesar 96% pada *dataset* mereka (60.802 citra). Sementara itu, Park *et al.* (2024) membandingkan metode deteksi *AI-generated image* serta menelaah aspek visualisasi dan generalisasi. Oleh karena itu, kebaruan penelitian ini bukan perbandingan CNN dan *Transformer* yang pertama. Aspek yang ditambahkan adalah pelaporan empat konfigurasi spesifik (ResNet50, MobileNetV2, ViT, dan model *Transformer* yang dilabeli DeiT) pada satu distribusi data wajah dan satu rangkaian metrik, disertai demonstrasi inferensi berbasis *Gradio*. Temuan studi sebelumnya tidak dapat dijadikan pembanding selisih *accuracy* secara langsung tanpa menyamakan *dataset*, generator, protokol *split*, dan skenario prapelatihan. Kontribusi tersebut bersifat deskriptif dan terbatas. Perbedaan bobot awal dan *preprocessing* belum terkendali sehingga eksperimen ini belum mengisolasi efek arsitektur. Empat konfigurasi dievaluasi menggunakan *dataset* citra wajah dan

E-ISSN: 2580-1643

metrik yang sama, yaitu *accuracy*, *precision*, *recall*, *F1-score*, serta *confusion matrix*. ResNet50 dan MobileNetV2 menggunakan inisialisasi ImageNet, sedangkan status bobot awal ViT dan *Transformer* berlabel DeiT belum dapat ditentukan dari konfigurasi yang tercatat. Dengan perbedaan ini, hasil penelitian diperlakukan sebagai perbandingan empiris antarimplementasi, bukan pengujian pengaruh arsitektur dalam kondisi pelatihan yang sepenuhnya setara. Pemilihan *checkpoint* dan independensi data uji juga menjadi faktor yang membatasi penafsiran hasil. *AI-generated image* merupakan citra digital yang dibentuk secara otomatis oleh model kecerdasan buatan generatif. Teknologi generatif dapat memanfaatkan pendekatan seperti *Generative Adversarial Network*, *Variational Autoencoder*, model difusi, maupun arsitektur berbasis *Transformer*. Perkembangan model tersebut memungkinkan sintesis wajah dengan bentuk, warna kulit, tekstur rambut, ekspresi, dan pencahayaan yang menyerupai fotografi. Kedekatan karakteristik visual antara karya manusia dan hasil AI membuat proses pembedaan semakin menantang dan membutuhkan dukungan metode komputasi (Subkhi *et al*., 2023). Pada kasus manipulasi visual yang lebih luas, kombinasi analisis artefak dan *deep learning* juga terbukti relevan untuk mendeteksi serta mengklasifikasikan citra *deepfake* (Rafique *et al*., 2023). Di sisi lain, *real image* dalam penelitian ini adalah citra wajah yang diperoleh melalui proses akuisisi menggunakan kamera atau perangkat optik.

Citra tersebut memiliki karakteristik yang dipengaruhi oleh kondisi lingkungan dan perangkat, seperti variasi cahaya, bayangan, *noise* sensor, *depth of field*, detail tekstur, serta artefak kompresi. Perbedaan antara citra asli dan sintetis tidak selalu berbentuk satu artefak yang konsisten sehingga model klasifikasi perlu mempelajari kombinasi pola lokal dan global agar dapat membedakan kedua kelas secara lebih *robust*. *Deep learning* memanfaatkan jaringan saraf bertingkat untuk mempelajari representasi data secara hierarkis. Pada citra, lapisan awal dapat menangkap pola sederhana seperti tepi, warna, dan tekstur, sedangkan lapisan yang lebih dalam menggabungkan informasi tersebut menjadi representasi objek yang lebih kompleks. Kinerja model dipengaruhi oleh kualitas dan jumlah data, *preprocessing*, arsitektur, serta strategi optimisasi. Dalam klasifikasi *AI-generated image*, *transfer learning* dan augmentasi menjadi pendekatan penting karena dapat memanfaatkan pengetahuan awal model sekaligus memperkaya variasi data latih. Pengujian terbaru juga menunjukkan bahwa *robustness* tetap perlu dinilai pada citra yang mengalami *noise* atau degradasi (Ranupadma *et al*., 2026). *Convolutional Neural Network* (CNN) membangun representasi citra melalui operasi konvolusi, fungsi aktivasi, *pooling*, dan lapisan klasifikasi. *Local receptive field* dan *parameter sharing* membuat CNN efisien dalam mengenali pola spasial yang berulang. ResNet50 memperkenalkan *residual learning* dan *shortcut connection* untuk mempertahankan aliran informasi pada jaringan yang dalam, sedangkan MobileNetV2 berfokus pada efisiensi melalui *depthwise separable convolution*, *inverted residual block*, dan *linear bottleneck*. Literatur terbaru memperlihatkan bahwa ResNet50 mampu menjadi ekstraktor fitur yang kuat untuk identifikasi konten visual AI dan *deepfake*, baik secara *end-to-end* maupun sebagai *feature extractor* yang dikombinasikan dengan *classifier* lain (Kusnaeni *et al*., 2025; Maulana *et al*., 2025; Rizqullah *et al*., 2026). *Vision Transformer* (ViT) mengadaptasi mekanisme *Transformer* untuk data visual dengan mengubah citra menjadi urutan *patch*. Setiap *patch* diproyeksikan ke ruang *embedding*, lalu diproses oleh *encoder* yang memanfaatkan *multi-head self-attention* dan *feed-forward network*. Mekanisme *self-attention* memungkinkan setiap *patch* berinteraksi dengan *patch* lain sehingga model dapat membangun representasi global. Keunggulan ini relevan ketika ciri pembeda antara citra asli dan *AI-generated* tidak hanya muncul pada satu area lokal. Namun, ViT cenderung membutuhkan data pelatihan yang besar agar representasinya berkembang secara optimal.

ViT mempelajari hubungan antarpotongan citra melalui *attention* (Dosovitskiy *et al*., 2020). Arsitektur DeiT dalam publikasi asalnya memperkenalkan pelatihan yang efisien terhadap data serta strategi *teacher–student* dengan *distillation token* (Touvron *et al*., 2021). Tidak semua *Transformer* berukuran kecil dapat disebut DeiT karena kesamaan jumlah blok, dimensi proyeksi, atau *attention heads* tidak cukup untuk mengidentifikasi distilasi. Oleh sebab itu, implementasi eksperimen ini disebut "model *Transformer* berlabel DeiT" sampai arsitektur dan kode pelatihannya diverifikasi. Evaluasi klasifikasi biner memerlukan metrik yang tidak hanya menunjukkan proporsi prediksi benar secara keseluruhan, tetapi juga keseimbangan kemampuan model dalam mengenali kelas. *Accuracy* menggambarkan proporsi klasifikasi yang benar, *precision* menilai ketepatan prediksi positif, *recall* mengukur kemampuan menangkap seluruh data positif, dan *F1-score* merangkum keseimbangan antara *precision* serta *recall*. *Confusion matrix* melengkapi metrik tersebut dengan memperlihatkan distribusi *True Positive*, *True Negative*, *False Positive*, dan *False Negative*. Penggunaan beberapa metrik sekaligus penting karena performa detektor *deepfake* dapat berubah antardataset dan skenario pengujian (Gong & Li, 2024; Rana *et al*., 2022). Setelah model dipilih,

E-ISSN: 2580-1643

deployment menghubungkan hasil eksperimen dengan antarmuka pengguna. *Gradio* menyediakan komponen antarmuka berbasis Python yang dapat menerima unggahan citra, memanggil fungsi inferensi, dan menampilkan hasil prediksi beserta *confidence score*. Pendekatan serupa telah digunakan untuk menyediakan antarmuka praktis pada sistem deteksi *deepfake* berbasis CNN (Gulo *et al.*, 2025). Dalam penelitian ini, *deployment* diposisikan sebagai prototipe untuk mendemonstrasikan pemanfaatan model terbaik, bukan sebagai *production environment* atau sistem autentikasi visual untuk keputusan berisiko tinggi.

Tabel 1. Ringkasan Penelitian Terkait

Studi	Model/Pendekatan	Fokus	Temuan Relevan
Subkhi *et al.* (2023)	CNN	Membedakan lukisan manusia dan AI	CNN mampu melakukan klasifikasi dengan baik; perbandingan dengan Transformer belum menjadi fokus.
Aria *et al.* (2025)	ViT dan EfficientNet-B0	Fake dan real image	Pendekatan berbasis ViT digunakan untuk klasifikasi gambar asli dan generatif.
Ranupadma *et al.* (2026)	CNN, transfer learning, augmentasi	AI-generated image	Augmentasi dan transfer learning meningkatkan robustness model CNN.
Erwin *et al.* (2025)	DeiT	Gambar AI dan manusia	DeiT menunjukkan performa kompetitif pada klasifikasi citra.
Park *et al.* (2024)	Beragam detektor dan ViT	Perbandingan deteksi AI-generated image	Performa dan generalisasi detektor berubah menurut karakteristik fitur, jenis generator, dan kondisi data uji.
Maulana *et al.* (2025)	ResNet50 dan fine-grained feature extraction	Identifikasi konten visual AI	ResNet50 dengan ekstraksi fitur granular efektif membedakan citra asli dan buatan AI.

Aria *et al.* (2025) melaporkan ViT 98% dan EfficientNet-B0 96% pada dataset berbeda; Park *et al.* (2024) juga telah menyajikan perbandingan detektor dan menekankan perubahan kinerja akibat karakteristik data. Karena kondisi uji dan generator penelitian-penelitian tersebut berbeda, perbandingan angka akurasi lintas artikel tanpa normalisasi protokol tidak sah. Celah yang ditangani studi ini terbatas pada penyajian hasil empat konfigurasi dalam satu pembagian data wajah yang dilaporkan, disertai confusion matrix dan demonstrasi alur Gradio. Belum ditangani: validasi lintas generator/dataset, audit kebocoran berbasis hash dan identitas, eksperimen inisialisasi yang seimbang, uji signifikansi prediksi berpasangan, serta pengukuran biaya komputasi. Dengan demikian kontribusi merupakan baseline empiris awal, bukan bukti superioritas intrinsik Transformer.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan eksperimen komparatif untuk mengevaluasi empat konfigurasi *deep learning* pada klasifikasi citra wajah asli dan sintetis. Seluruh model dinilai menggunakan pembagian data dan metrik yang sama. ResNet50 dan MobileNetV2 menggunakan *transfer learning* dari ImageNet, sedangkan asal bobot awal kedua *Transformer* tidak tercatat secara terperinci. Perbedaan inisialisasi dan kebutuhan *preprocessing* tersebut membatasi interpretasi perbedaan akurasi sebagai akibat arsitektur semata. Alur penelitian mengikuti *Cross Industry Standard Process for Data Mining* (CRISP-DM) yang terdiri atas *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Pada tahap *Business Understanding*, ditetapkan masalah utama berupa semakin sulitnya membedakan citra wajah *AI-generated* dan *real image*. Tahap *Data Understanding* digunakan untuk memahami struktur, distribusi, format, dan karakteristik visual *dataset*. *Data Preparation* menyiapkan citra agar memiliki format yang konsisten. Tahap *Modeling* menerapkan ResNet50, MobileNetV2, ViT, dan DeiT, sedangkan *Evaluation* mengukur performa pada data uji. Model terbaik kemudian dibawa ke tahap *Deployment* melalui aplikasi *Gradio*. *Dataset* penelitian berasal dari koleksi publik *140k Real and Fake Faces* yang diunggah oleh xhlulu pada Kaggle. Sebanyak 28.000 citra digunakan, terdiri atas 14.000 citra asli dan 14.000 citra sintetis. Pembagiannya mencakup 20.000 citra pelatihan, 4.000 citra validasi, dan 4.000 citra pengujian. Prosedur penarikan sampel dari keseluruhan koleksi, identitas generator sintetis, ketentuan lisensi, kriteria eksklusi, serta *seed* pembagian tidak tercatat secara lengkap. Pemeriksaan *hash* berkas, kemiripan visual, dan kesamaan identitas wajah antarsubset juga tidak menghasilkan laporan audit yang dapat dianalisis. Oleh karena itu, keterpisahan

E-ISSN: 2580-1643

subset secara administratif belum cukup untuk memastikan tidak adanya kebocoran data. Seluruh citra diubah ukurannya dan dinormalisasi menggunakan *rescale* 1/255. Augmentasi pada data latih meliputi rotasi hingga 25°, *horizontal flip*, zoom 0,2, serta pergeseran horizontal dan vertikal sebesar 0,1, sedangkan data validasi dan pengujian tidak diaugmentasi. Penerapan skema normalisasi tunggal ini merupakan karakteristik eksperimen yang dilaksanakan. Karena bobot ImageNet pada ResNet50 dan MobileNetV2 memiliki kebutuhan prapemrosesan yang berbeda, skema tersebut berpotensi memengaruhi kinerja kedua model. Oleh sebab itu, hasil ini tidak mengisolasi pengaruh arsitektur dari kesesuaian masukan. Pada implementasi berbasis Keras, *preprocessing* ImageNet untuk ResNet50 melibatkan perubahan urutan kanal RGB menjadi BGR dan pemusatan terhadap rerata ImageNet, sedangkan MobileNetV2 lazim menggunakan transformasi masukan ke rentang −1 hingga 1. Kebutuhan *preprocessing* ViT dan *Transformer* berlabel DeiT bergantung pada definisi model serta bobot yang digunakan. Dalam eksperimen ini, keempat model tetap diproses menggunakan *rescale* 1/255. Oleh karena itu, perbedaan kinerja perlu ditafsirkan berdasarkan pilihan prapemrosesan tersebut. Eksperimen lanjutan dengan transformasi yang sesuai dan skenario inisialisasi yang sebanding diperlukan untuk menilai ketahanan hasil.

Tabel 2. Distribusi Dataset

Subset	Real Image	AI-generated Image	Total
Training	10.000	10.000	20.000
Validation	2.000	2.000	4.000
Testing	2.000	2.000	4.000
Total	14.000	14.000	28.000

ResNet50 dan MobileNetV2 menggunakan bobot awal ImageNet dengan *base model* yang dibekukan. Pada ResNet50, lapisan klasifikasi diganti dengan *Global Average Pooling 2D*, *Batch Normalization*, *Dense*, *Dropout*, dan keluaran *sigmoid*, sedangkan MobileNetV2 menggunakan *classifier* biner di atas *base model*. Pada kedua *Transformer*, riwayat bobot awal tidak tercatat secara rinci. Oleh sebab itu, selisih performa yang diamati merepresentasikan kombinasi arsitektur, inisialisasi, dan konfigurasi pemrosesan, bukan efek arsitektur secara terpisah. ViT menggunakan masukan 224×224 piksel, *patch* 16×16, dimensi proyeksi 64, empat *attention heads*, dan delapan lapisan *Transformer*. Model *Transformer* berlabel DeiT menggunakan masukan 224×224 piksel, *patch* 16×16, dimensi proyeksi 96, enam *attention heads*, dan 12 lapisan. Informasi implementasi yang tersedia tidak mencakup *distillation token*, model *teacher*, suhu distilasi, bobot *distillation loss*, maupun *checkpoint* pralatih. Oleh karena itu, istilah "*Transformer* berlabel DeiT" digunakan untuk hasil eksperimen ini, sedangkan mekanisme distilasi sebagaimana diperkenalkan Touvron *et al.* (2021) belum dapat dipastikan berlangsung. Pelatihan menggunakan *optimizer* Adam, *binary cross-entropy*, *batch size* 32, dan maksimum 10 *epoch*. *Callback* yang digunakan meliputi *EarlyStopping*, *ReduceLROnPlateau*, dan *ModelCheckpoint*. Nilai *learning rate* awal dan minimum, β_1, β_2, *epsilon*, metrik monitor, *patience*, *min_delta*, *factor*, *seed*, ukuran *Dense*, rasio *Dropout*, ambang klasifikasi, serta *checkpoint* akhir tidak terdokumentasi secara lengkap. Tabel 4 merinci parameter yang dapat dilaporkan dan parameter yang belum diketahui. Ketiadaan rincian tersebut membatasi reprodusibilitas dan independensi pemilihan model terbaik.

Tabel 3. Ringkasan Konfigurasi Model

Model	Kelompok	Konfigurasi utama	Karakteristik
ResNet50	CNN	Pre-trained ImageNet; base model dibekukan; GAP2D, BatchNorm, Dense-ReLU, Dropout, sigmoid	Transfer learning
MobileNetV2	CNN	Pre-trained ImageNet; base model dibekukan; classifier biner	Depthwise separable convolution dan inverted residual
ViT	Transformer	224×224; patch 16×16; 196 patch; projection 64; 4 heads; 8 Transformer layers	Self-attention

E-ISSN: 2580-1643

DeiT	Transformer	224×224; patch 16×16; projection 96; 6 heads; 12 Transformer layers	Self-attention; data-efficient Transformer

Tabel 4. Keterlaporan konfigurasi pelatihan empat model

Komponen	Informasi yang dapat diverifikasi	Informasi yang masih diperlukan
Inisialisasi	ResNet50 dan MobileNetV2: ImageNet; base dibekukan	ViT dan Transformer berlabel DeiT: scratch atau bobot pralatih, sumber checkpoint
Input / augmentasi	Resize; rescale 1/255; rotasi 25°, flip horizontal, zoom 0,2, shift 0,1	Preprocessing khusus bobot untuk percobaan pembanding
Optimisasi	Adam; binary crossentropy	learning rate awal dan minimum; beta_1, beta_2, epsilon
Pelatihan	Batch size 32; maksimum 10 epoch	random seed untuk split dan inisialisasi; learning-rate schedule aktual
Callbacks	EarlyStopping, ReduceLROnPlateau, ModelCheckpoint disebut	monitor, patience, factor, min_delta, restore_best_weights, path/epoch checkpoint
Lapisan keluaran	Klasifikasi biner / sigmoid pada konfigurasi CNN	Ukuran Dense, Dropout, ambang keputusan dan spesifikasi akhir masing-masing model
Pemilihan model	Hasil performa uji tersedia	Aturan checkpoint berdasarkan validasi dan bukti test set terkunci

Evaluasi klasifikasi menggunakan 4.000 citra uji yang ditempatkan terpisah dari subset pelatihan dan validasi. Performa dinilai menggunakan accuracy, precision, recall, F1-score, dan confusion matrix. Akan tetapi, tidak tersedia hasil pemeriksaan identitas dan duplikasi citra lintas subset untuk mengonfirmasi independensinya pada tingkat konten. Selain itu, aturan pemilihan checkpoint berdasarkan metrik validasi tidak tercatat secara lengkap. Keterbatasan tersebut diperhitungkan dalam interpretasi hasil pengujian. Secara matematis, metrik evaluasi dinyatakan sebagai berikut.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$
$$Precision = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}$$
$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

Model dengan hasil evaluasi terbaik dipilih sebagai final model untuk deployment. Model disimpan agar dapat dimuat kembali tanpa pelatihan ulang, kemudian diintegrasikan ke aplikasi Gradio. Antarmuka menerima unggahan citra, menjalankan preprocessing yang konsisten dengan tahap pelatihan, melakukan inferensi, dan menampilkan kategori Real Image atau AI-generated Image beserta confidence score.

3. Hasil dan Pembahasan

3.1 Hasil

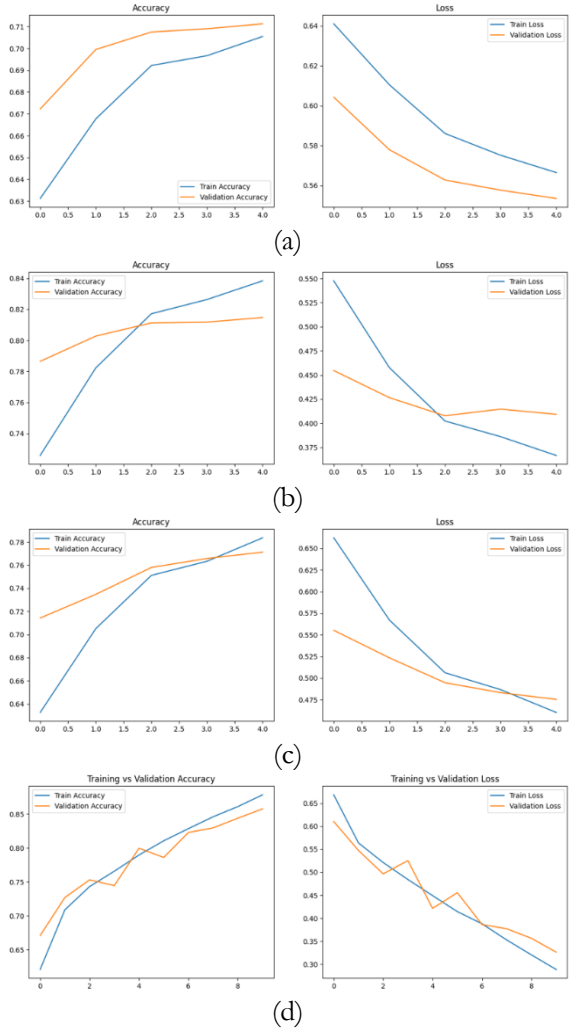
Dinamika Pelatihan Model

Gambar 1 memperlihatkan kurva pelatihan dan validasi untuk ResNet50, MobileNetV2, ViT, serta Transformer berlabel DeiT. Keempat panel memuat *training accuracy*, *validation accuracy*, *training loss*, dan *validation loss* sepanjang *epoch* yang tercatat. Kurva ResNet50, MobileNetV2, dan ViT mencakup lima *epoch*, sedangkan kurva Transformer berlabel DeiT mencakup sepuluh *epoch*. Penyebab perbedaan durasi pelatihan tidak dapat ditetapkan hanya dari grafik karena riwayat penghentian dini tidak tersedia secara lengkap. Berdasarkan ringkasan metrik akhir pada Tabel 5 dan kurva asli pada Gambar 1, ResNet50 menghasilkan *training accuracy* 70,6% dan *validation accuracy* 71,1% (*validation loss* 0,554), MobileNetV2 menghasilkan 83,8% dan 81,5% (0,409), ViT menghasilkan 78,3% dan 77,1% (0,475), sedangkan model Transformer berlabel DeiT menghasilkan 87,8% dan 85,8% (0,327). Angka tersebut merupakan metrik pelatihan dan validasi, bukan *accuracy* pengujian akhir, sehingga performa uji ditampilkan secara terpisah. Penilaian *overfitting* tidak dapat disimpulkan hanya dari satu nilai akhir.

E-ISSN: 2580-1643

Tabel 5. Ringkasan Tren Pelatihan

Model	Training Accuracy (%)	Validation Accuracy (%)	Training Loss	Validation Loss
ResNet50	63,1 → 70,6	67,2 → 71,1	0,641 → 0,567	0,604 → 0,554
MobileNetV2	72,6 → 83,8	78,7 → 81,5	0,548 → 0,367	0,454 → 0,409
DeiT	62,0 → 87,8	67,0 → 85,8	0,668 → 0,289	0,610 → 0,327
ViT	63,3 → 78,3	71,4 → 77,1	0,662 → 0,461	0,555 → 0,475



Gambar 1. Perbandingan dinamika pelatihan empat model klasifikasi: (a) ResNet50, (b) MobileNetV2, (c) Vision Transformer (ViT), dan (d) model Transformer berlabel DeiT berdasarkan nilai loss dan accuracy pada data pelatihan serta validasi setiap epoch

Gambar 1 menyajikan dinamika pelatihan empat model, yaitu ResNet50, MobileNetV2, *Vision Transformer* (ViT), dan model *Transformer* berlabel DeiT. Setiap panel memperlihatkan perubahan nilai *training loss*, *validation loss*, *training accuracy*, dan *validation accuracy* pada setiap *epoch*. Visualisasi ini digunakan untuk mengamati perkembangan kemampuan model dalam mempelajari karakteristik citra, tingkat konvergensi selama pelatihan, serta perbedaan perilaku optimisasi antara model berbasis CNN dan *Transformer*. Evaluasi kurva pelatihan dilakukan dengan memperhatikan arah perubahan *loss* dan *accuracy* serta kesenjangan antara data pelatihan dan validasi. Penurunan *loss* yang disertai peningkatan *accuracy* pada kedua subset dapat menunjukkan bahwa proses pembelajaran berlangsung dengan baik. Sebaliknya, penurunan *training loss* yang tidak diikuti perbaikan *validation loss* dapat menjadi indikasi *overfitting*. Nilai *accuracy* yang relatif rendah pada kedua subset dapat mengindikasikan *underfitting*, meskipun kesimpulan tersebut perlu mempertimbangkan kompleksitas model, konfigurasi pelatihan, dan jumlah *epoch* yang digunakan. Batas pelatihan adalah 10 *epoch*, sedangkan panjang kurva aktual berbeda antarmodel. ResNet50, MobileNetV2, dan ViT memperlihatkan lima *epoch*, sementara *Transformer* berlabel DeiT memperlihatkan sepuluh *epoch*. Secara umum, kurva menunjukkan penurunan *loss* dan peningkatan *accuracy*, meskipun terdapat fluktuasi pada sejumlah nilai validasi. Grafik mendukung pemeriksaan awal proses konvergensi, tetapi tidak memastikan bahwa seluruh model mencapai titik optimum atau menjalani kondisi pelatihan yang identik.

Hasil Pengujian Setiap Model

Pada ResNet50, *confusion matrix* menunjukkan 1.349 citra *Fake* dan 1.324 citra *Real* yang diklasifikasikan dengan benar. Kesalahan terdiri atas 651 citra *Fake* yang diprediksi sebagai *Real* dan 676 citra *Real* yang diprediksi sebagai *Fake*. Dari nilai tersebut diperoleh *accuracy* 66,83%, *precision* 66,62%, *recall* 67,45%, dan *F1-score* 67,03%. *Classification report* menampilkan nilai sekitar 0,67 untuk *precision*, *recall*, dan *F1-score* pada kedua kelas, yang menunjukkan bahwa kinerja ResNet50 relatif seimbang, namun masih memiliki tingkat kesalahan klasifikasi yang cukup besar. MobileNetV2 menghasilkan 1.469 prediksi benar pada kelas *Fake* dan 1.648 prediksi benar pada kelas *Real*. Sebanyak 531 citra *Fake* salah diklasifikasikan sebagai *Real*, dan 352 citra *Real* salah diklasifikasikan sebagai *Fake*. *Accuracy* model mencapai 77,93%, *precision* 80,67%, *recall* 73,45%, dan *F1-score* 76,89%. Kelas *Fake* memperoleh *precision* 0,81, *recall* 0,73, dan *F1-score* 0,77, sedangkan kelas *Real* memperoleh *precision* 0,76, *recall* 0,82, dan *F1-score* 0,79. Hasil ini memperlihatkan peningkatan yang jelas dibandingkan ResNet50. DeiT memberikan jumlah prediksi benar tertinggi, yaitu 1.744 citra *Fake* dan 1.735 citra *Real*. Kesalahan klasifikasi hanya terdiri atas 256 citra *Fake* yang diprediksi *Real* dan 265 citra *Real* yang diprediksi *Fake*. *Accuracy* mencapai 86,98%, *precision* 86,81%, *recall*

E-ISSN: 2580-1643

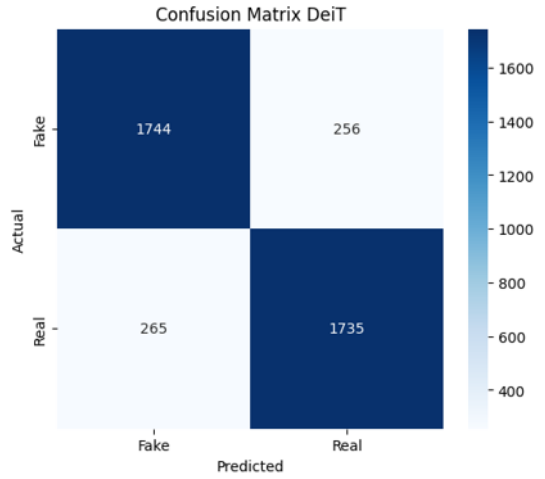
87,20%, dan *F1-score* 87,00%. *Classification report* menghasilkan *precision*, *recall*, dan *F1-score* sekitar 0,87 pada kedua kelas, yang menunjukkan performa tinggi sekaligus seimbang. ViT mengklasifikasikan 1.421 citra *Fake* dan 1.452 citra *Real* secara benar, sedangkan 579 citra *Fake* diprediksi sebagai *Real* dan 548 citra *Real* diprediksi sebagai *Fake*. Model memperoleh *accuracy* 71,83%, *precision* 72,17%, *recall* 71,05%, dan *F1-score* 71,60%. Nilai per kelas berada di sekitar 0,71–0,73. Oleh karena itu, ViT mampu melakukan klasifikasi dengan cukup baik, tetapi performanya masih berada di bawah DeiT dan MobileNetV2 pada konfigurasi penelitian ini.

Tabel 6. Hasil Evaluasi Pada Data Pengujian

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet50	66,83	66,62	67,45	67,03
MobileNetV2	77,93	80,67	73,45	76,89
ViT	71,83	72,17	71,05	71,60
DeiT	86,98	86,81	87,20	87,00

Tabel 7. Ringkasan Confusion Matrix

Model	Fake→Fake	Fake→Real	Real→Fake	Real→Real
ResNet50	1.349	651	676	1.324
MobileNetV2	1.469	531	352	1.648
ViT	1.421	579	548	1.452
DeiT	1.744	256	265	1.735



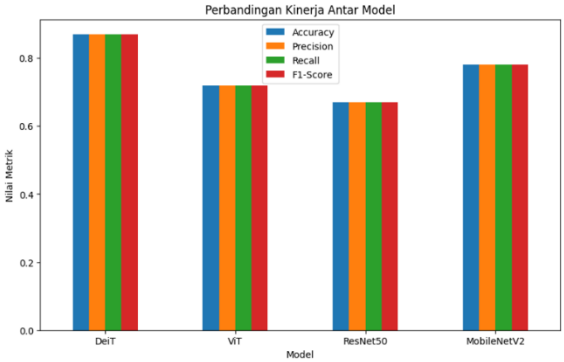
Gambar 2. Confusion Matrix Model DeiT

Perbandingan CNN dan Transformer

Rata-rata tanpa pembulatan awal dihitung menggunakan jumlah prediksi benar dari *confusion matrix*: CNN = ((2.673/4.000) + (3.117/4.000))/2 × 100% = 72,375% ≈ 72,38%, dan Transformer = ((2.873/4.000) + (3.479/4.000))/2 × 100% = 79,400% ≈ 79,40%. Selisih nilai yang belum dibulatkan adalah 7,025 poin persentase, yang dibulatkan dua desimal menjadi 7,03 poin persentase. Perbandingan ini bersifat deskriptif karena bobot awal, *preprocessing*, pemilihan *checkpoint*, dan audit independensi data belum setara atau belum tervalidasi. Rata-rata *accuracy* kelompok CNN dihitung dari ResNet50 dan MobileNetV2, sedangkan kelompok *Transformer* dihitung dari ViT dan *Transformer* berlabel DeiT. Hasil perhitungan berdasarkan angka sebelum pembulatan menunjukkan rata-rata CNN sebesar 72,38% dan *Transformer* sebesar 79,40%, atau berbeda 7,03 poin persentase. Selisih tersebut merupakan perbedaan absolut, bukan kenaikan relatif sebesar 7%. Karena inisialisasi dan *preprocessing* belum sepenuhnya sebanding, temuan ini tidak menunjukkan superioritas intrinsik satu keluarga arsitektur. Kinerja MobileNetV2 yang melampaui ViT turut memperlihatkan peran konfigurasi model. ResNet50 memperoleh performa terendah meskipun *residual learning* membantu pelatihan jaringan yang dalam. Pada kasus ini, ekstraksi fitur lokal berbasis konvolusi belum cukup untuk menangkap seluruh pola pembeda antara gambar *AI-generated* dan gambar asli. Ciri sintesis dapat tersebar pada tekstur, rambut, pencahayaan, dan latar belakang, sehingga hubungan antarbagian citra yang lebih luas dapat menjadi informasi penting. Temuan bahwa ResNet tetap mampu mendeteksi konten visual berbasis AI konsisten dengan Maulana *et al.* (2025), Kusnaeni *et al.* (2025), dan Rizqullah *et al.* (2026), tetapi perbandingan langsung pada penelitian ini menunjukkan bahwa arsitektur lain dapat memberikan performa yang lebih tinggi pada *dataset* yang digunakan. *MobileNetV2* mengungguli ResNet50 sebesar sekitar 11 poin persentase pada *accuracy*. *Depthwise separable convolution* dan *inverted residual* memungkinkan model mempertahankan informasi penting dengan struktur yang lebih efisien. Hasil ini memperlihatkan bahwa arsitektur yang lebih ringan tidak selalu menghasilkan akurasi yang lebih rendah. Pada eksperimen ini, efisiensi representasi MobileNetV2 justru memberikan generalisasi yang lebih baik daripada ResNet50. Model *Transformer* yang dilaporkan sebagai DeiT memperoleh *accuracy* 86,98%, *precision* 86,81%, *recall* 87,20%, dan *F1-score* 87,00% pada data uji yang dilaporkan. Hasil tersebut merupakan skor terbaik

E-ISSN: 2580-1643

di antara empat konfigurasi yang diuji, tetapi tidak membuktikan bahwa mekanisme hubungan global, distilasi, atau efisiensi penggunaan data menjadi penyebabnya. Penjelasan mekanistik tersebut masih berupa hipotesis karena tidak tersedia *ablation study*, validasi *distillation token* atau *teacher*, maupun hasil pengukuran FLOPs, parameter, memori, dan waktu inferensi. Secara keseluruhan, DeiT memperoleh hasil terbaik pada konfigurasi yang dilaporkan, dengan *accuracy* 86,98%, *precision* 86,81%, *recall* 87,20%, dan *F1-score* 87,00%. Hasil ini menunjukkan performa empiris yang lebih tinggi pada *dataset* pengujian yang digunakan, tetapi belum cukup untuk menyimpulkan bahwa keunggulan tersebut disebabkan oleh *self-attention*, efisiensi data, atau distilasi pengetahuan. Kesimpulan mekanistik memerlukan verifikasi implementasi DeiT dan, idealnya, eksperimen *ablation* serta pengukuran biaya komputasi. Karena rincian *distillation token* dan *teacher* belum dilaporkan, interpretasi tentang distilasi harus diperlakukan sebagai belum terverifikasi. Hasil ini konsisten dengan Erwin *et al.* (2025) yang menunjukkan efektivitas DeiT untuk membedakan gambar AI dan manusia, serta mendukung temuan bahwa pendekatan *patch-wise Transformer* memiliki potensi tinggi pada klasifikasi citra sintetis (Arshed *et al.*, 2024).



Gambar 3. Perbandingan Kinerja Seluruh Model

Model dengan akurasi pengujian tertinggi disimpan dalam format .keras dan diintegrasikan dengan antarmuka *Gradio*. Pemilihan ini didasarkan pada peringkat *accuracy* yang tercatat dalam eksperimen, dan belum dapat dipastikan berasal dari prosedur seleksi *checkpoint* berbasis validasi yang independen. Oleh karena mekanisme distilasi belum terkonfirmasi, model *deployment* dirujuk sebagai *Transformer* berlabel DeiT.

Implementasi Model Terbaik pada Gradio

Aplikasi *Gradio* menerima unggahan citra, melakukan *preprocessing* dan inferensi, kemudian menampilkan prediksi kelas beserta *confidence score*. Empat contoh pengujian pada Tabel 8 memperlihatkan keberhasilan alur inferensi untuk sampel tersebut. Prosedur pemilihan sampel, jumlah pengulangan, spesifikasi perangkat, *warm-up*, dan distribusi latensi belum tercatat, sehingga hasil *deployment* terbatas pada evaluasi fungsional. Pada pengujian fungsional *Gradio*, empat citra masukan menghasilkan kelas prediksi yang sesuai dengan label aktual. Pengujian ini menunjukkan bahwa alur penerimaan citra, *preprocessing*, inferensi, dan penampilan keluaran dapat berfungsi pada contoh yang digunakan. Jumlah sampel yang kecil dan tidak adanya protokol pemilihan terstandar membatasi generalisasi temuan, sedangkan waktu inferensi juga tidak diukur pada tahap ini.

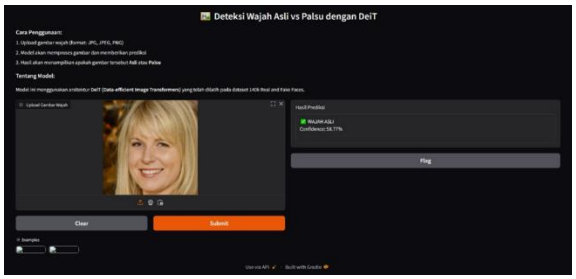
Tabel 8. Hasil Pengujian Deployment

No.	Label aktual	Hasil prediksi	Confidence
1	Real Image	Real Image	58,77%
2	AI-generated Image	AI-generated Image	94,14%
3	AI-generated Image	AI-generated Image	96,72%
4	Real Image	Real Image	76,77%

Evaluasi latensi bukan bagian dari hasil pengujian yang tersedia. Pengukuran lanjutan dapat menggunakan sampel acak dari *test set* independen dengan *seed* dan identitas berkas yang dicatat, disertai *warm-up* dan pengulangan per sampel. Komponen waktu yang relevan meliputi *preprocessing*, *forward pass*, dan *end-to-end* (ms/citra), dengan ringkasan median dan p95 serta spesifikasi CPU/GPU. Karena pengukuran tersebut belum dilakukan, penelitian ini tidak melaporkan angka latensi. Pengujian *deployment* dilakukan menggunakan empat citra, yang terdiri atas dua citra *Real Image* dan dua citra *AI-generated Image*. Seluruh citra memperoleh hasil prediksi yang sesuai dengan label aktual, dengan *confidence* masing-masing sebesar 58,77%, 94,14%, 96,72%, dan 76,77%. Hasil tersebut menunjukkan bahwa prototipe berbasis *Gradio* dapat menjalankan proses penerimaan gambar, *preprocessing*, inferensi, dan penyajian hasil klasifikasi pada sampel yang

E-ISSN: 2580-1643

diuji. Evaluasi *deployment* dalam penelitian ini terbatas pada verifikasi fungsional empat contoh citra. *Seed* dan asal sampel dari *test set* terkunci belum dapat ditetapkan, sedangkan pencatatan *warm-up*, durasi *preprocessing*, *forward pass*, *end-to-end*, median, p95, dan spesifikasi komputasi belum tersedia. Oleh karena itu, hasil *deployment* tidak menjadi dasar klaim kecepatan atau efisiensi komputasi model.



Gambar 4. Hasil Prediksi Real Image Pada Prototipe Gradio

Kinerja terbaik model *Transformer* berlabel DeiT pada konfigurasi ini tidak otomatis menunjukkan efisiensi komputasi ataupun dampak kausal perhatian global. *MobileNetV2* adalah model dengan *accuracy* kedua tertinggi dalam data yang dilaporkan, tetapi klaim lebih ringan, lebih cepat, atau lebih hemat energi belum diuji secara empiris pada perangkat yang sama. Kajian lanjutan perlu melaporkan jumlah parameter, FLOPs, *peak memory*, waktu latih, dan latensi inferensi yang diukur menggunakan protokol yang seragam.
Keterbatasan utama meliputi satu *dataset* citra wajah, 10 *epoch* maksimum, *sampling* dan lisensi *dataset* yang belum diverifikasi, tidak adanya audit duplikasi serta identitas wajah lintas *split*, bobot awal *Transformer* dan implementasi distilasi yang belum terdokumentasi, *preprocessing* yang berbeda dari kontrak bobot pralatih, dan potensi bias seleksi dari data uji. Belum tersedia pengulangan beberapa *random seed*, prediksi individual berpasangan untuk uji *McNemar*, interval kepercayaan, pengujian lintas generator, maupun pengukuran waktu dan memori. Uji *McNemar* harus dilakukan atas prediksi benar atau salah dua model pada citra uji yang sama, bukan atas ringkasan *accuracy* atau *confusion matrix* agregat. Kesimpulan dibatasi sebagai pengamatan satu eksperimen, dan sistem *Gradio* belum dapat diklaim siap digunakan untuk autentikasi berisiko tinggi.

Implikasi dan Keterbatasan
Hasil penelitian memperlihatkan bahwa pemilihan arsitektur untuk deteksi *AI-generated image* perlu mempertimbangkan karakteristik data dan efisiensi pembelajaran, bukan hanya kedalaman atau popularitas model. *DeiT* menjadi pilihan paling kuat pada eksperimen ini, tetapi *MobileNetV2* juga menarik karena mampu mencapai performa kedua terbaik dengan karakteristik arsitektur yang lebih ringan. Bagi sistem yang memprioritaskan keterbatasan sumber daya komputasi, temuan tersebut menunjukkan adanya *trade-off* yang perlu dipertimbangkan antara performa maksimum dan efisiensi model. Generalisasi temuan dibatasi oleh penggunaan satu *dataset* citra wajah, total 28.000 citra terpilih, dan pelatihan maksimum 10 *epoch*. Audit duplikasi dan identitas lintas *subset*, pengujian lintas *dataset*, pengulangan dengan beberapa *random seed*, interval kepercayaan, uji *McNemar*, serta pengukuran sumber daya komputasi belum dilakukan atau belum terdokumentasi. Hasil diperlakukan sebagai perbandingan empiris pada konfigurasi yang diteliti. Studi selanjutnya dapat menggunakan *preprocessing* sesuai bobot, protokol inisialisasi yang sepadan, evaluasi lintas generator, dan pengukuran latensi. Sejalan dengan literatur, pergeseran generator, kompresi, *noise*, dan domain citra merupakan aspek penting dalam pengujian ketahanan detektor (Bammey, 2023; Dong *et al.*, 2023; Manisha *et al.*, 2025; Park *et al.*, 2024; Ranupadma *et al.*, 2026).

3.2 Pembahasan
Hasil pengujian menunjukkan urutan performa yang konsisten pada keempat konfigurasi, yaitu DeiT (*accuracy* 86,98%), MobileNetV2 (77,93%), ViT (71,83%), dan ResNet50 (66,83%). Temuan bahwa model berbasis Transformer dapat mencapai performa kompetitif dalam membedakan citra buatan AI dan manusia sejalan dengan Erwin *et al.* (2025), yang melaporkan efektivitas DeiT pada tugas serupa. Pendekatan berbasis *patch* dan Transformer juga dilaporkan efektif untuk klasifikasi *deepfake* multikelas (Arshed *et al.*, 2024), dan Aria *et al.* (2025) menunjukkan bahwa ViT dapat membedakan citra asli dan generatif. Namun, tingkat *accuracy* yang dicapai dalam penelitian ini lebih rendah daripada angka 98% yang dilaporkan Aria *et al.* (2025). Perbedaan tersebut wajar mengingat dataset, generator, dan protokol pembagian data yang berbeda. Oleh karena itu, angka lintas penelitian tidak dapat dibandingkan secara langsung. Apabila hasil dikelompokkan berdasarkan arsitektur, rata-rata *accuracy* CNN (ResNet50 dan MobileNetV2) sebesar 72,38%, sedangkan rata-rata kelompok Transformer (ViT dan DeiT) sebesar 79,40%, dengan selisih 7,03 poin persentase. Selisih ini bersifat deskriptif dan tidak dapat diartikan sebagai keunggulan intrinsik keluarga arsitektur tertentu. Bobot awal, *preprocessing*, pemilihan *checkpoint*, dan audit independensi data belum setara atau belum tervalidasi. Kinerja MobileNetV2 yang lebih tinggi dibandingkan ViT juga menunjukkan bahwa konfigurasi model turut menentukan hasil. Temuan ini sejalan dengan Park *et al.* (2024), yang menekankan bahwa performa detektor dapat

E-ISSN: 2580-1643

berubah menurut karakteristik fitur dan kondisi data uji. ResNet50 memperoleh performa terendah dalam penelitian ini. Meskipun *residual learning* membantu pelatihan jaringan yang dalam, ekstraksi fitur lokal berbasis konvolusi tampaknya belum cukup untuk menangkap seluruh pola pembeda. Ciri sintesis dapat tersebar pada tekstur rambut, pencahayaan, dan latar belakang, sehingga hubungan antarbagian citra yang lebih luas menjadi informasi penting. Hasil ini berbeda dengan sejumlah studi yang menunjukkan kemampuan ResNet yang baik dalam mengidentifikasi konten visual AI, seperti Maulana *et al.* (2025), Kusnaeni *et al.* (2025), dan Rizqullah *et al.* (2026). Perbedaan tersebut mengindikasikan bahwa kemampuan ResNet50 sangat bergantung pada dataset, konfigurasi pelatihan, dan strategi ekstraksi fitur yang digunakan. Selain itu, MobileNetV2 mengungguli ResNet50 sekitar 11 poin persentase pada *accuracy*. Keunggulan ini dapat dijelaskan oleh *depthwise separable convolution* dan *inverted residual* yang memungkinkan model mempertahankan informasi penting dengan struktur yang lebih efisien. Hasil ini menunjukkan bahwa arsitektur yang lebih ringan tidak selalu menghasilkan akurasi yang lebih rendah. Temuan ini juga sejalan dengan pandangan bahwa pemilihan arsitektur sebaiknya mempertimbangkan karakteristik data dan efisiensi pembelajaran, bukan hanya kedalaman model. Meski demikian, belum ada studi pembanding dalam tinjauan pustaka yang menguji hubungan ini secara langsung, sehingga temuan tersebut perlu dikonfirmasi pada penelitian lanjutan.

Model Transformer berlabel DeiT memperoleh skor tertinggi pada seluruh metrik, dengan *precision* 86,81%, *recall* 87,20%, dan *F1-score* 87,00%. Keseimbangan antara *precision* dan *recall* ini menunjukkan bahwa model tidak hanya unggul pada satu kelas tertentu. Secara mekanistik, *self-attention* pada Vision Transformer memungkinkan setiap *patch* berinteraksi dengan *patch* lainnya sehingga model dapat membangun representasi global (Dosovitskiy *et al.*, 2020). Mekanisme ini relevan ketika ciri pembeda antara citra asli dan sintetis tersebar pada beberapa area. Namun, penjelasan tersebut masih merupakan hipotesis. Penelitian ini tidak menyediakan *ablation study*, verifikasi *distillation token* atau *teacher* sebagaimana diperkenalkan Touvron *et al.* (2021), maupun pengukuran FLOPs, parameter, dan memori. Dengan demikian, keunggulan DeiT belum dapat diatribusikan pada *self-attention*, efisiensi data, atau distilasi pengetahuan. Dari sisi implementasi, prototipe Gradio yang dibangun pada penelitian ini memiliki pola yang serupa dengan sistem deteksi *deepfake* berbasis CNN yang dikembangkan Gulo *et al.* (2025). Keduanya menyediakan antarmuka sederhana untuk menerima unggahan citra dan menampilkan hasil prediksi beserta *confidence score*. Pada pengujian fungsional, empat citra menghasilkan prediksi yang sesuai dengan label aktual. Namun, jumlah sampel yang kecil dan tidak adanya protokol pemilihan yang terstandar membatasi generalisasi temuan tersebut. Hasil ini hanya membuktikan bahwa alur penerimaan citra, *preprocessing*, inferensi, dan penyajian keluaran dapat berjalan dengan baik. Hasil ini belum dapat dijadikan dasar klaim kecepatan atau efisiensi komputasi. Keterbatasan penelitian ini juga perlu dipertimbangkan secara serius. Penggunaan satu dataset, pelatihan maksimum 10 *epoch*, serta ketiadaan audit duplikasi dan identitas wajah lintas subset membuat keterpisahan data belum dapat dipastikan. Hal ini sejalan dengan kekhawatiran Gong dan Li (2024) serta Rana *et al.* (2022) mengenai pentingnya pengujian yang komprehensif pada detektor *deepfake*. Selain itu, Bammey (2023) dan Manisha *et al.* (2025) menegaskan bahwa generalisasi detektor terhadap generator dan pascaproses citra yang berbeda merupakan persoalan utama. Sejalan dengan literatur tersebut, pengujian lintas generator, kompresi, dan domain citra perlu dilakukan sebelum hasil ini dapat digeneralisasi. Pengujian statistik seperti uji McNemar pada prediksi berpasangan dan pengulangan dengan beberapa *random seed* juga diperlukan untuk memperkuat kesimpulan. Secara keseluruhan, hasil penelitian memberikan implikasi bahwa pemilihan arsitektur untuk deteksi citra AI-generated perlu mempertimbangkan karakteristik data dan kebutuhan komputasi. DeiT menjadi pilihan paling kuat pada eksperimen ini, sedangkan MobileNetV2 menawarkan trade-off yang menarik bagi sistem yang membatasi sumber daya komputasi. Meski demikian, penelitian ini masih berupa baseline empiris awal. Kontribusinya terletak pada penyajian empat konfigurasi dalam satu distribusi data wajah yang dilaporkan, disertai *confusion matrix* dan demonstrasi Gradio, bukan pada pembuktian superioritas intrinsik Transformer atas CNN.

4. Kesimpulan

Berdasarkan hasil *confusion matrix* sebelum pembulatan, rerata *accuracy* CNN adalah 72,38% dan rerata Transformer adalah 79,40%, dengan selisih 7,03 poin persentase. Perbedaan ini tidak dapat diatribusikan semata-mata pada arsitektur, karena strategi prapelatihan dan *preprocessing* belum dikendalikan, serta belum tersedia uji statistik maupun pengulangan eksperimen. Identitas implementasi DeiT juga masih memerlukan verifikasi kode distilasi.

Perbandingan empat konfigurasi model menunjukkan perbedaan kinerja dalam klasifikasi

E-ISSN: 2580-1643

citra *AI-generated* dan *real image*. MobileNetV2 memperoleh *accuracy* 77,93%, lebih tinggi dibandingkan ResNet50 (66,83%) dan ViT (71,83%). Model yang dilaporkan sebagai DeiT memperoleh nilai tertinggi, yaitu *accuracy* 86,98%, *precision* 86,81%, *recall* 87,20%, dan *F1-score* 87,00%. Interpretasi hasil perlu dibatasi pada dataset dan konfigurasi yang digunakan, terutama karena rincian prapelatihan dan mekanisme distilasi Transformer belum lengkap. Perhitungan menggunakan nilai sebelum pembulatan menghasilkan rata-rata *accuracy* CNN sebesar 72,38% dan Transformer sebesar 79,40%, dengan selisih 7,03 poin persentase. Temuan ini bersifat deskriptif dan berlaku pada konfigurasi serta dataset yang digunakan. Pengaruh khusus arsitektur belum dapat dipisahkan dari variasi bobot awal, *preprocessing*, dan hiperparameter pelatihan. Model dengan hasil pengujian tertinggi telah diimplementasikan sebagai prototipe Gradio yang menerima citra dan mengembalikan label serta *confidence score*. Empat contoh pengujian menunjukkan fungsi dasar aplikasi, tetapi evaluasi latensi dan pemilihan sampel secara terstandar belum tersedia. Keterbatasan lain mencakup rincian hiperparameter, pemeriksaan kebocoran data, identitas mekanisme distilasi, pengulangan eksperimen dengan *seed* berbeda, pengujian berpasangan, dan evaluasi lintas dataset. Dengan demikian, implementasi ini merupakan bukti konsep klasifikasi interaktif, bukan sistem autentikasi siap produksi.

5. Daftar Pustaka

Aria, M. S. A., Slamet, C., & Firdaus, M. D. (2025). Klasifikasi fake dan real menggunakan Vision Transformer dan EfficientNet-B0 pada gambar asli dan generatif AI. *SMATIKA Jurnal, 15*(01), 179–192. https://doi.org/10.32664/smatika.v15i01.1531.

Arshed, M. A., Alwadain, A., Faizan Ali, R., Mumtaz, S., Ibrahim, M., & Muneer, A. (2023). Unmasking deception: Empowering deepfake detection with vision transformer network. *Mathematics, 11*(17), Article 3710. https://doi.org/10.3390/math11173710.

Arshed, M. A., Mumtaz, S., Ibrahim, M., Dewi, C., Tanveer, M., & Ahmed, S. (2024). Multiclass AI-generated deepfake face detection using patch-wise deep learning model. *Computers, 13*(1), Article 31. https://doi.org/10.3390/computers13010031.

Bammey, Q. (2023). Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing, 5*, 1–9. https://doi.org/10.1109/OJSP.2023.3337714.

Dong, F., Zou, X., Wang, J., & Liu, X. (2023). Contrastive learning-based general deepfake detection with multi-scale RGB frequency clues. *Journal of King Saud University: Computer and Information Sciences, 35*(4), 90–99. https://doi.org/10.1016/j.jksuci.2023.03.005.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.11929.

Erwin, I. T., Arda, A. L., Taufik, I., & Erwin, H. A. (2025). Implementasi model DeiT untuk membedakan gambar buatan AI dan manusia pada ilustrasi animasi 2D. *INTI Nusa Mandiri, 19*(2), 172–180. https://doi.org/10.33480/inti.v19i2.6306.

Gong, L. Y., & Li, X. J. (2024). A contemporary survey on deepfake detection: Datasets, algorithms, and challenges. *Electronics, 13*(3), Article 585. https://doi.org/10.3390/electronics13030585.

Gulo, S. A., Pertiwi, A. A., Nasution, S. P. S., & Syahputra, H. (2025). Deteksi deepfake dalam citra menggunakan Convolutional Neural Network (CNN). *JATI (Jurnal Mahasiswa Teknik Informatika), 9*(5), 8655–8660. https://doi.org/10.36040/jati.v9i5.14896.

Kusnaeni, K., Adriani, I. R., Hafid, M. S., & Rizal, M. E. (2025). Deepfake image classification using ResNet50 feature extraction and XGBoost learning model. *Journal of Mathematics, Computations and Statistics, 8*(2), 258–268. https://doi.org/10.35580/jmathcos.v8i2.8387.

Malik, A., Kuribayashi, M., Abdullahi, S. M., & Khan, A. N. (2022). DeepFake detection

E-ISSN: 2580-1643

for human face images and videos: A survey. *IEEE Access, 10*, 18757–18775. https://doi.org/10.1109/ACCESS.2022.3151186.

Manisha, M., Li, C.-T., & Kotegar, K. A. (2025). AI-synthesized image detection: Source camera fingerprinting to discern the authenticity of digital images. *IEEE Access, 13*, 29660–29672. https://doi.org/10.1109/ACCESS.2025.3537558.

Maulana, D., Azzahrah, N. S., Syach, R. A., Syauri, S., Syafi'i, M., & Rozikin, C. (2025). Identifikasi konten visual buatan AI dengan ResNet dan fine-grained feature extraction. *Jurnal Informatika dan Teknik Elektro Terapan, 13*(3S1), 1248–1256. https://doi.org/10.23960/jitet.v13i3S1.8056.

Park, D., Na, H., & Choi, D. (2024). Performance comparison and visualization of AI-generated-image detection methods. *IEEE Access, 12*, 62609–62627. https://doi.org/10.1109/ACCESS.2024.3394250.

Rafique, R., Gantassi, R., Amin, R., Frnda, J., Mustapha, A., & Alshehri, A. H. (2023). Deep fake detection and classification using error-level analysis and deep learning. *Scientific Reports, 13*(1), Article 7422. https://doi.org/10.1038/s41598-023-34629-3.

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access, 10*, 25494–25513. https://doi.org/10.1109/ACCESS.2022.3154404.

Ranupadma, N. K. R., Dharma, E. M., & Paramitha, A. A. I. I. (2026). Implementasi transfer learning dan augmentasi pada model CNN untuk meningkatkan robustness klasifikasi AI-generated image. *JATI (Jurnal Mahasiswa Teknik Informatika), 10*(2), 2370–2377. https://doi.org/10.36040/jati.v10i2.17412.

Rizqullah, G. F., Eosina, P., & Pramuko, A. E. K. (2026). Enhanced performance evaluation of VGG16 and ResNet50 for deepfake detection using local ternary pattern. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 10*(1), 701–712. https://doi.org/10.33395/sinkron.v10i1.15582.

Sihananto, A. N., Al Haromainy, M. M., Fauzi, Z. A., Reza, R. A., Putra, G. C. H., & Christianty, T. M. (2024). Wayang's images recognition using Vision Transformer. *IJDASEA (International Journal of Data Science, Engineering, and Analytics), 4*(2), 15–27. https://doi.org/10.33005/ijdasea.v4i2.24.

Subkhi, M. B., Setiawan, A. B., & Candra, M. Y. A. (2023). Klasifikasi gambar: Membedakan lukisan buatan manusia dan AI dengan CNN. *Paradigma: Jurnal Filsafat, Sains, Teknologi, dan Sosial Budaya, 29*(4), 149–155. https://doi.org/10.33503/paradigma.v29i4.1284.

Tahyudin, G. G., Sulistiyo, M. D., Arzaki, M., & Rachmawati, E. (2024). Classifying gender based on face images using vision transformer. *JOIV: International Journal on Informatics Visualization, 8*(1), 18–25. https://doi.org/10.62527/joiv.8.1.1923.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 10347–10357). PMLR.