

Perbandingan EfficientNetB0, MobileNetV2, dan DenseNet121 Berbasis Transfer Learning untuk Deteksi AI-Generated Images

Comparison of EfficientNetB0, MobileNetV2, and DenseNet121 Based on Transfer Learning for AI-Generated Image Detection

Nicholas Abraham Herytra, Suprayogi*

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia

Email: ¹111202214811@mhs.dinus.ac.id, ^{2,*}suprayogi@dsn.dinus.ac.id

Email Penulis Korespondensi: suprayogi@dsn.dinus.ac.id

Received: 08/07/2026, Revised: 21/07/2026, Accepted: 29/09/2026, Published: 30/09/2026

Abstrak—Perkembangan model generatif berbasis kecerdasan buatan (AI) seperti GAN dan Diffusion Model telah menghasilkan gambar sintesis berkualitas tinggi yang sulit dibedakan secara visual dari gambar nyata. Hal ini menimbulkan tantangan serius dalam bidang keamanan digital, verifikasi media, dan kepercayaan publik terhadap konten visual. Namun, masih sedikit penelitian yang membandingkan beberapa arsitektur CNN dengan strategi *two-phase transfer learning* pada dataset yang sama. Penelitian ini mengusulkan pendekatan berbasis transfer learning dengan membandingkan tiga arsitektur Convolutional Neural Network (CNN) pretrained ImageNet, yaitu EfficientNetB0, MobileNetV2, dan DenseNet121, untuk mendeteksi gambar yang dihasilkan oleh AI. Dataset yang digunakan adalah AI vs Real Images Dataset dari Kaggle yang terdiri dari dua kelas: *AI-generated images* dan gambar nyata. Pelatihan dilakukan dalam dua fase: fase pertama dengan backbone yang dibekukan untuk mengadaptasi bobot pretrained, dan fase kedua berupa fine-tuning pada 30 layer terakhir dari backbone. Augmentasi data mencakup rotasi, pergeseran, zoom, shear, dan perubahan kecerahan diterapkan pada data latih, serta class weighting dan label smoothing digunakan untuk menangani ketidakseimbangan kelas. Evaluasi dilakukan menggunakan metrik akurasi, presisi, recall, F1-score, dan AUC-ROC. Hasil eksperimen menunjukkan bahwa DenseNet121 mencapai performa tertinggi dengan akurasi sebesar 98,00%, presisi 99,10%, recall 98,21%, F1-score 98,65%, dan AUC-ROC 99,72%. EfficientNetB0 menempati posisi kedua dengan akurasi 95,33%, sementara MobileNetV2 mencatat akurasi 94,67% dengan jumlah parameter paling sedikit sehingga unggul dari sisi efisiensi. Visualisasi Grad-CAM mengonfirmasi bahwa model memusatkan perhatian pada objek utama citra serta wilayah bertekstur dan bertransisi warna yang menjadi penanda artefak generatif. Penelitian ini menunjukkan bahwa pendekatan transfer learning dua fase dengan augmentasi data yang terstruktur efektif untuk tugas deteksi gambar AI.

Kata Kunci: Deteksi Gambar AI; Transfer Learning; EfficientNetB0; MobileNetV2; DenseNet121; Convolutional Neural Network

Abstract—The rapid advancement of AI-based generative models such as GANs and Diffusion Models has produced high-quality synthetic images that are visually indistinguishable from real photographs. This poses serious challenges in digital security, media verification, and public trust in visual content. However, few studies have compared multiple CNN architectures using a two-phase transfer learning strategy on the same dataset. This study proposes a transfer learning-based approach by comparing three ImageNet-pretrained CNN architectures, namely EfficientNetB0, MobileNetV2, and DenseNet121, for detecting AI-generated images. The dataset used is the AI vs Real Images Dataset from Kaggle, consisting of two classes: AI-generated and real images. Training was conducted in two phases: the first phase with a frozen backbone to adapt pretrained weights, and the second phase involving fine-tuning the last 30 layers of the backbone. Data augmentation including rotation, shifting, zoom, shear, and brightness variation was applied to the training set, while class weighting and label smoothing were employed to address class imbalance. Evaluation was performed using accuracy, precision, recall, F1-score, and AUC-ROC metrics. Experimental results show that DenseNet121 achieved the highest performance with an accuracy of 98.00%, precision 99.10%, recall 98.21%, F1-score 98.65%, and AUC-ROC 99.72%. EfficientNetB0 ranked second with 95.33% accuracy, while MobileNetV2 recorded 94.67% accuracy with the fewest parameters, making it superior in terms of efficiency. Grad-CAM visualizations confirm that the models focus on the main objects as well as textured and color-transition regions that serve as markers of generative artifacts. This study demonstrates that a two-phase transfer learning approach with structured data augmentation is effective for AI image detection tasks.

Keywords: AI Image Detection; Transfer Learning; EfficientNetB0; MobileNetV2; DenseNet121; Convolutional Neural Network

1. PENDAHULUAN

Perkembangan teknologi kecerdasan buatan (AI), khususnya model generatif berbasis deep learning seperti Generative Adversarial Network (GAN) dan Diffusion Model, telah memungkinkan pembuatan gambar sintesis dengan kualitas yang sangat tinggi [1]. Gambar-gambar yang dihasilkan oleh model tersebut sering kali sulit dibedakan secara visual dari gambar fotografis nyata, bahkan oleh manusia yang terlatih sekalipun [2]. Fenomena ini menimbulkan ancaman serius dalam berbagai domain, mulai dari penyebaran berita palsu (disinformasi) [3], manipulasi media sosial, pembuatan deepfake [4], [5], hingga pemalsuan identitas dan penipuan digital.

Kemampuan model generatif modern untuk menghasilkan wajah manusia, pemandangan, dan objek yang sangat realistis telah mendorong kebutuhan mendesak terhadap sistem deteksi otomatis yang andal. Deteksi gambar AI (AI-generated image detection) menjadi bidang penelitian yang semakin penting, terutama di era di mana produksi konten visual sintesis menjadi semakin mudah dan murah dengan alat-alat seperti Midjourney, DALL-E, dan Stable Diffusion [6]. Kemudahan akses terhadap alat-alat generatif ini menyebabkan volume gambar sintesis yang beredar di

internet meningkat secara eksponensial, sehingga mekanisme verifikasi keaslian gambar menjadi komponen kritis dalam ekosistem informasi digital yang sehat.

Dampak negatif dari penyalahgunaan *AI-generated images* telah terdokumentasi secara luas dalam berbagai kasus nyata [7]. Di bidang politik, gambar deepfake tokoh publik digunakan untuk menyebarkan propaganda dan menurunkan kredibilitas individu atau institusi tertentu [4]. Di sektor keuangan, gambar identitas palsu yang dihasilkan oleh AI dimanfaatkan untuk melewati sistem verifikasi Know Your Customer (KYC) pada platform perbankan digital. Di ranah hukum, manipulasi bukti visual menjadi ancaman terhadap integritas proses peradilan. Sementara itu, di media sosial, gambar hoaks yang tampak autentik mempercepat penyebaran narasi yang menyesatkan publik. Realitas ini menegaskan bahwa sistem deteksi gambar AI bukan sekadar persoalan teknis semata, melainkan telah menjadi kebutuhan mendasar dalam menjaga kepercayaan terhadap informasi visual di ruang publik digital.

Urgensi permasalahan ini semakin nyata dalam beberapa tahun terakhir, ketika gambar sintesis tidak hanya digunakan untuk hiburan, tetapi juga menjadi instrumen dalam kampanye disinformasi berskala besar. Berbagai kajian menunjukkan bahwa konten visual yang dimanipulasi memiliki daya persuasi yang tinggi karena manusia cenderung lebih memercayai bukti visual dibandingkan teks [3]. Kondisi ini diperburuk oleh fakta bahwa kemampuan masyarakat untuk mengenali gambar palsu masih terbatas, khususnya di wilayah dengan literasi media yang rendah. Oleh karena itu, ketersediaan sistem deteksi otomatis yang akurat dan dapat dipercaya menjadi salah satu garis pertahanan utama dalam menjaga integritas ekosistem informasi modern.

Dari sudut pandang teknis, deteksi gambar AI merupakan permasalahan klasifikasi biner yang menantang karena *AI-generated images* modern tidak lagi mengandung artefak visual yang kasat mata seperti pada generasi GAN awal [8], [9]. Model-model generatif terkini menghasilkan gambar yang secara statistik sangat mirip dengan distribusi gambar nyata, sehingga metode berbasis aturan (*rule-based*) dan analisis piksel sederhana tidak lagi memadai. Pendekatan berbasis *deep learning*, khususnya *Convolutional Neural Network (CNN)*, telah menunjukkan kemampuan superior dalam mengekstrak fitur representatif pada level hierarkis yang tidak dapat ditangkap oleh metode konvensional [10]. Namun demikian, tantangan utama tetap ada dalam hal generalisasi model terhadap gambar dari generator yang belum pernah dilihat selama pelatihan [11], serta ketahanan model terhadap berbagai jenis *postprocessing* seperti kompresi JPEG, *resizing*, dan penambahan *noise* [12].

Berbagai pendekatan telah diusulkan untuk mendeteksi gambar yang dihasilkan oleh AI [13]. Penelitian awal banyak berfokus pada analisis artefak frekuensi tinggi yang merupakan jejak khas proses generasi [6]. Beberapa studi mendemonstrasikan bahwa *classifier CNN* yang dilatih pada satu jenis generator dapat digeneralisasi ke generator lain melalui augmentasi data yang tepat [10]. Pendekatan berbasis frekuensi seperti analisis spektrum Fourier juga telah terbukti efektif dalam mengidentifikasi pola statistik yang membedakan gambar AI dari gambar nyata, khususnya pada gambar hasil *diffusion model* [6]. Selain itu, penelitian berbasis dataset benchmark berskala besar seperti *CIFAKE* menunjukkan bahwa arsitektur CNN mampu mengklasifikasikan gambar nyata dan sintesis dengan akurasi tinggi sekaligus menyediakan interpretabilitas melalui visualisasi [14]. Pendekatan terbaru bahkan mengembangkan metode purifikasi fitur artefak untuk meningkatkan kemampuan deteksi lintas domain (*cross-domain*) [12].

Selain pendekatan berbasis CNN murni, sejumlah penelitian terbaru mengembangkan arsitektur hibrida dan mekanisme atensi untuk meningkatkan akurasi deteksi. Gunukula et al. [15] menggabungkan arsitektur residual dengan blok *Squeeze-and-Excitation (SE)* untuk memberikan bobot adaptif pada fitur yang paling diskriminatif, sementara Hasaniyyah et al. [16] menunjukkan bahwa integrasi mekanisme SE ke dalam *DenseNet-121* mampu meningkatkan performa klasifikasi gambar nyata dan AI pada dataset benchmark. Pendekatan berbasis *graph attention network* juga diusulkan untuk menangkap relasi spasial antar wilayah gambar [17]. Perkembangan ini menegaskan bahwa pemilihan arsitektur backbone dan strategi ekstraksi fitur memiliki pengaruh signifikan terhadap keberhasilan tugas deteksi, sehingga perbandingan sistematis antar arsitektur menjadi relevan untuk dilakukan.

Transfer learning telah terbukti sebagai strategi yang efektif dalam berbagai tugas klasifikasi gambar dengan keterbatasan data, karena memungkinkan model memanfaatkan representasi fitur visual yang telah dipelajari dari dataset berskala besar seperti *ImageNet* [18]. Dalam konteks deteksi gambar AI, *transfer learning* menawarkan keuntungan ganda: pertama, model pretrained telah memiliki pemahaman mendalam tentang fitur visual dasar seperti tepi, tekstur, dan bentuk yang relevan untuk mendeteksi inkonsistensi pada *AI-generated images*; kedua, kebutuhan data pelatihan berlabel yang lebih sedikit dibandingkan pelatihan dari awal (*training from scratch*) sangat relevan mengingat kelangkaan dataset deteksi gambar AI yang besar dan terdiversifikasi. Strategi *fine-tuning* selektif, di mana hanya sebagian layer terakhir dari backbone yang diaktifkan untuk diperbarui bobotnya, memungkinkan adaptasi fitur pretrained ke domain target tanpa mengorbankan pengetahuan umum yang telah diperoleh dari *pretraining* [18].

Tiga arsitektur CNN yang menjadi fokus perbandingan dalam penelitian ini dipilih berdasarkan pertimbangan representativitas dan relevansi praktis. *EfficientNetB0* mewakili generasi terbaru arsitektur CNN dengan efisiensi komputasi tinggi melalui *compound scaling*. *MobileNetV2* mewakili arsitektur ringan yang dirancang untuk deployment pada perangkat mobile dan *edge computing* dengan keterbatasan memori dan daya komputasi. *DenseNet121* mewakili paradigma *dense connectivity* yang mendorong penggunaan ulang fitur secara maksimal. Ketiga arsitektur ini telah banyak digunakan sebagai backbone pada berbagai studi klasifikasi citra, termasuk perbandingan *transfer learning* pada domain lain [18], sehingga perbandingan sistematis di antara ketiganya akan memberikan wawasan yang komprehensif bagi praktisi yang perlu memilih arsitektur sesuai dengan kebutuhan dan kendala sumber daya komputasi yang tersedia.

Di samping aspek akurasi, kebutuhan akan interpretabilitas model juga semakin mendapat perhatian dalam penelitian deteksi gambar AI. Sistem deteksi yang bersifat black-box sulit dipercaya dalam aplikasi sensitif seperti forensik digital dan verifikasi media, karena pengguna tidak dapat memahami dasar pengambilan keputusan model [19]. Teknik explainable AI (XAI), khususnya visualisasi berbasis gradien, memungkinkan peneliti untuk mengidentifikasi wilayah gambar yang menjadi fokus perhatian model saat melakukan klasifikasi. Integrasi teknik interpretabilitas ini ke dalam alur evaluasi tidak hanya meningkatkan transparansi, tetapi juga membantu memverifikasi bahwa model benar-benar mempelajari fitur yang relevan secara forensik, bukan sekadar bias dataset.

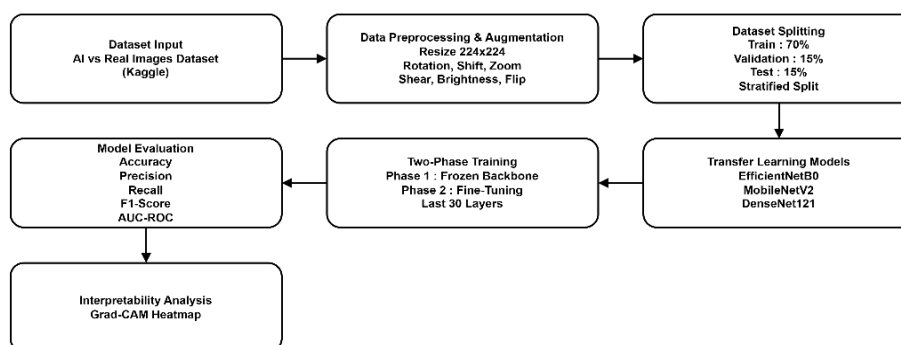
Meskipun berbagai pendekatan deteksi gambar AI telah dikembangkan [7], [10], terdapat tiga celah yang belum terjawab secara memadai dalam literatur. *Pertama*, sebagian besar studi hanya mengevaluasi satu arsitektur CNN tanpa perbandingan sistematis antar arsitektur pada dataset dan kondisi eksperimen yang identik, sehingga sulit menentukan arsitektur yang paling sesuai untuk kebutuhan praktis tertentu. *Kedua*, efektivitas strategi *two-phase transfer learning* yang menggabungkan fase *frozen backbone* dengan fase *fine-tuning* selektif masih jarang dievaluasi secara eksplisit, padahal strategi ini berpotensi menyeimbangkan adaptasi domain dan efisiensi komputasi. *Ketiga*, aspek interpretabilitas melalui teknik visualisasi seperti Grad-CAM belum banyak diintegrasikan sebagai bagian dari evaluasi, sehingga verifikasi bahwa model mempelajari fitur yang relevan secara forensik, bukan sekadar bias dataset, masih terbatas [19]. Ketiga celah inilah yang menjadi motivasi utama penelitian ini.

Berdasarkan celah tersebut, penelitian ini bertujuan membandingkan performa tiga arsitektur CNN *pretrained*, yaitu EfficientNetB0, MobileNetV2, dan DenseNet121, dalam mendeteksi gambar yang dihasilkan AI menggunakan pendekatan *transfer learning* dua fase [18]. Ketiga arsitektur dipilih sebagai *metode pembandingan* karena mewakili tiga paradigma desain berbeda: *compound scaling* (EfficientNetB0), arsitektur ringan untuk *edge deployment* (MobileNetV2), dan *dense connectivity* (DenseNet121). Kontribusi utama penelitian ini meliputi: (1) analisis komparatif sistematis ketiga arsitektur pada dataset yang sama dengan kondisi eksperimen identik; (2) implementasi strategi pelatihan dua fase dengan augmentasi data terstruktur pada dataset *imbalanced*; (3) evaluasi komprehensif menggunakan lima metrik kinerja (akurasi, presisi, *recall*, F1-score, dan AUC-ROC); serta (4) analisis interpretabilitas melalui visualisasi Grad-CAM untuk memahami dasar keputusan model. Hasil penelitian diharapkan memberikan panduan praktis berbasis bukti empiris dalam memilih arsitektur yang optimal sesuai kebutuhan aplikasi.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilaksanakan melalui serangkaian tahapan yang sistematis dan terstruktur. Tahapan penelitian meliputi: (1) pengumpulan dan eksplorasi dataset, (2) preprocessing dan augmentasi data, (3) pembagian dataset, (4) pembangunan model transfer learning, (5) pelatihan dua fase, (6) evaluasi dan perbandingan model, serta (7) visualisasi interpretabilitas menggunakan Grad-CAM. Seluruh eksperimen dilakukan menggunakan bahasa pemrograman Python 3.10 dengan framework TensorFlow 2.x di lingkungan Kaggle Notebooks yang dilengkapi GPU akselerator NVIDIA Tesla T4. Pendekatan bertahap semacam ini sejalan dengan praktik umum pada penelitian deteksi gambar AI berbasis deep learning yang menekankan reproduktibilitas dan evaluasi yang menyeluruh [1], [10]. Untuk memberikan gambaran menyeluruh mengenai ruang lingkup dan urutan proses penelitian, keseluruhan tahapan divisualisasikan dalam bentuk diagram alur pada Gambar 1.



Gambar 1. Pipeline tahapan penelitian deteksi gambar AI

2.2 Objek Penelitian

Objek penelitian ini adalah tugas deteksi gambar AI (AI-generated image detection) yang diformulasikan sebagai masalah klasifikasi biner antara citra hasil generasi model AI dan citra fotografis nyata. Pendekatan yang diteliti adalah transfer learning dua fase dengan memanfaatkan tiga arsitektur Convolutional Neural Network (CNN) pretrained ImageNet sebagai objek pembandingan, yaitu EfficientNetB0, MobileNetV2, dan DenseNet121. Ketiga arsitektur ini dievaluasi pada dataset yang identik untuk memperoleh perbandingan yang adil (*fair comparison*). Ruang lingkup analisis mencakup empat aspek: (1) karakteristik dan distribusi data yang digunakan, (2) strategi pelatihan dua fase

(frozen backbone dan fine-tuning selektif), (3) kinerja klasifikasi berdasarkan lima metrik evaluasi, dan (4) interpretabilitas model melalui visualisasi Grad-CAM. Dengan demikian, objek penelitian mencakup baik aspek arsitektur model maupun karakteristik data yang dianalisis, sebagaimana dirangkum dalam kata kunci penelitian.

2.3 Dataset

Dataset yang digunakan dalam penelitian ini adalah AI vs Real Images Dataset yang bersumber dari platform Kaggle (rhythmghai/ai-vs-real-images-dataset). Dataset ini terdiri dari dua kelas: gambar yang dihasilkan oleh model AI generatif (kelas AI) dan gambar fotografis nyata (kelas Real). Dataset mencakup berbagai kategori visual termasuk wajah manusia, pemandangan alam, objek sehari-hari, dan arsitektur, sehingga memberikan keragaman yang memadai untuk pelatihan model yang mampu menggeneralisasi berbagai jenis konten visual. Penggunaan dataset yang mencakup beragam kategori visual penting untuk menghindari bias model terhadap domain tertentu, sebagaimana ditekankan dalam studi deteksi lintas domain [12].

Pemilihan AI vs Real Images Dataset didasarkan pada beberapa pertimbangan. Pertama, dataset ini mencakup beragam kategori visual dunia nyata seperti wajah manusia, pemandangan, makanan, dan arsitektur, berbeda dengan dataset seperti CIFAKE yang berbasis citra CIFAR-10 beresolusi rendah (32×32 piksel), sehingga lebih merepresentasikan skenario deteksi pada citra beresolusi tinggi. Kedua, citra AI pada dataset ini dihasilkan oleh beragam generator modern, sehingga relevan dengan kondisi terkini. Ketiga, ukuran dataset yang moderat sesuai dengan keterbatasan sumber daya komputasi pada lingkungan Kaggle Notebooks, sekaligus memungkinkan penerapan strategi transfer learning yang justru dirancang untuk kondisi data terbatas. Meskipun demikian, penggunaan dataset benchmark berskala lebih besar seperti GenImage atau Synthbuster tetap direkomendasikan untuk penelitian selanjutnya guna menguji generalisasi model secara lebih menyeluruh.

Distribusi kelas pada dataset bersifat tidak seimbang (imbalanced), dengan kelas Real (74,87%) mendominasi dibandingkan kelas AI-Generated (25,13%). Kondisi ini mencerminkan skenario dunia nyata di mana gambar nyata jauh lebih banyak beredar dibandingkan gambar sintetis. Untuk menangani ketidakseimbangan ini, class weight dihitung secara otomatis menggunakan metode balanced weighting. Seluruh gambar diubah ukurannya menjadi 224×224 piksel sesuai persyaratan input arsitektur CNN yang digunakan. Format gambar yang diterima meliputi JPEG, JPG, dan PNG.

Tabel 1. Statistik Dataset

Kelas	Jumlah Gambar	Persentase (%)
AI-Generated	250	25,13
Real	745	74,87
Total	995	100,00

2.4 Preprocessing dan Augmentasi Data

Preprocessing data dilakukan secara berbeda untuk setiap arsitektur model sesuai dengan fungsi normalisasi yang direkomendasikan oleh masing-masing arsitektur pretrained. EfficientNetB0 menggunakan fungsi preprocess_input dari modul efficientnet yang menskalakan nilai piksel ke rentang yang sesuai, MobileNetV2 menggunakan normalisasi mobilenet_v2 yang menskalakan ke $[-1, 1]$, sedangkan DenseNet121 menggunakan normalisasi densenet yang mengikuti konvensi ImageNet. Augmentasi data diterapkan hanya pada data pelatihan (training set) untuk meningkatkan variasi dan mencegah overfitting. Teknik augmentasi yang digunakan meliputi: rotasi gambar hingga 25 derajat, pergeseran horizontal dan vertikal hingga 20%, pembesaran (zoom) hingga 20%, shear transformation sebesar 0,15, variasi kecerahan dalam rentang $[0.8, 1.2]$, dan pembalikan horizontal (horizontal flip). Mode pengisian piksel yang hilang akibat transformasi geometri menggunakan metode nearest neighbor. Strategi augmentasi yang terstruktur seperti ini terbukti meningkatkan kemampuan generalisasi model klasifikasi citra berbasis transfer learning, khususnya pada dataset berukuran terbatas [18].

2.5 Pembagian Dataset

Dataset dibagi menjadi tiga subset menggunakan stratified split untuk memastikan distribusi kelas yang proporsional pada setiap subset. Pembagian yang diterapkan adalah 70% untuk data pelatihan (training set), 15% untuk data validasi (validation set), dan 15% untuk data pengujian (test set). Stratified splitting dilakukan menggunakan fungsi train_test_split dari scikit-learn dengan parameter random_state=42 untuk reproducibility.

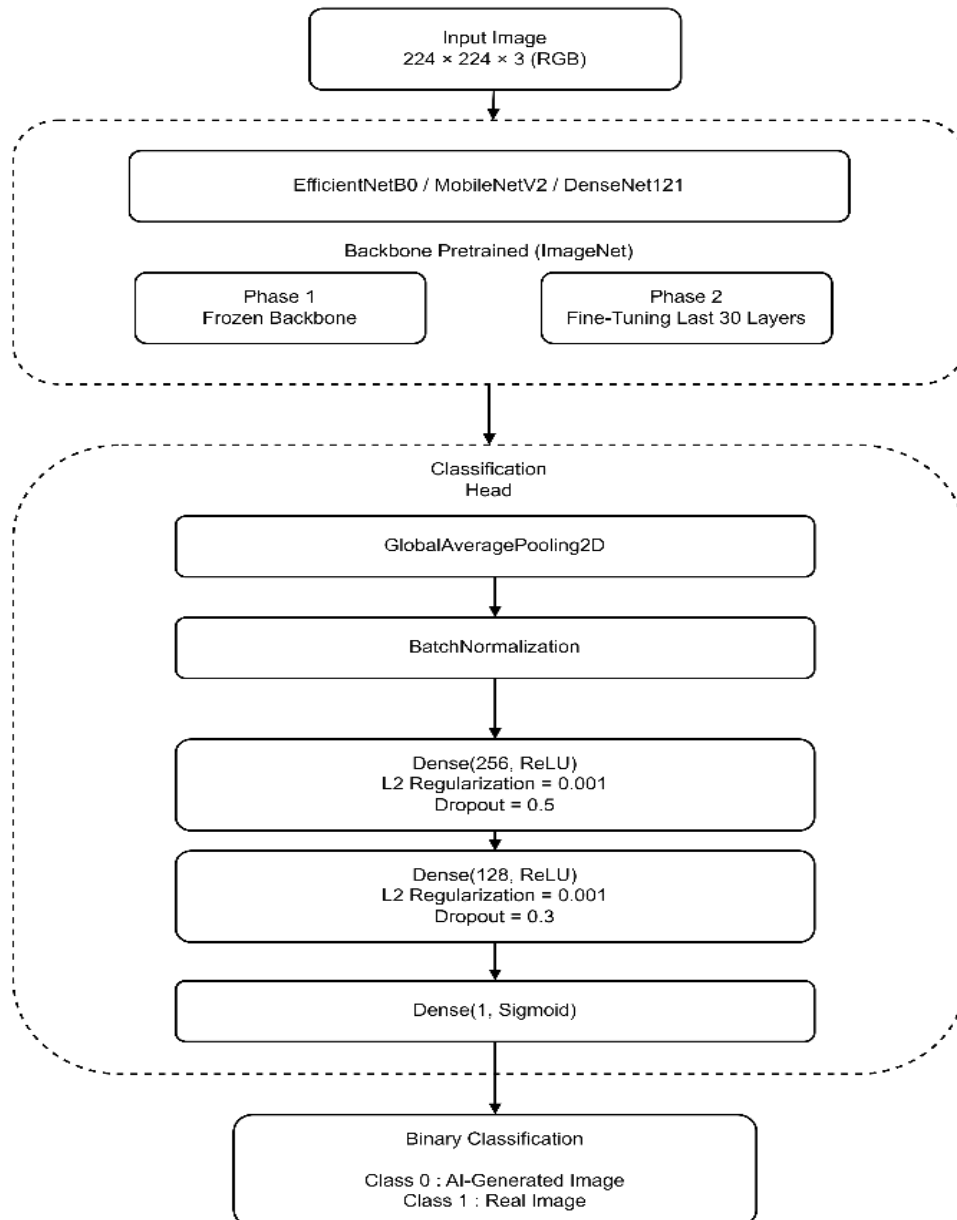
Tabel 2. Pembagian Dataset

Subset	Jumlah Gambar	Persentase (%)
Training	696	70
Validation	149	15
Test	150	15

Data uji sebanyak 150 citra terdiri atas 38 citra AI-Generated dan 112 citra Real, mengikuti proporsi kelas hasil stratified split.

2.6 Arsitektur Model

Penelitian ini membandingkan tiga arsitektur CNN yang telah pretrained pada dataset ImageNet, yaitu EfficientNetB0, MobileNetV2, dan DenseNet121. Ketiga model memiliki karakteristik arsitektur yang berbeda dan dikenal efektif dalam tugas klasifikasi gambar, termasuk pada deteksi gambar AI [15], [16]. Gambaran umum arsitektur model transfer learning yang digunakan, mulai dari input citra hingga lapisan output, ditunjukkan pada Gambar 2.



Gambar 2. Arsitektur model transfer learning dengan custom classification head

EfficientNetB0 merupakan arsitektur CNN yang memanfaatkan compound scaling untuk menyeimbangkan kedalaman, lebar, dan resolusi jaringan secara simultan. EfficientNetB0 sebagai varian terkecil dalam keluarga EfficientNet memiliki sekitar 5,3 juta parameter dan mencapai trade-off terbaik antara akurasi dan efisiensi komputasi, sehingga banyak digunakan sebagai backbone transfer learning pada berbagai tugas klasifikasi citra [18].

MobileNetV2 menggunakan blok inverted residual dan linear bottleneck yang dirancang untuk deployment pada perangkat dengan sumber daya komputasi terbatas. MobileNetV2 memiliki sekitar 3,4 juta parameter, menjadikannya model paling ringan di antara ketiga arsitektur yang dibandingkan, dan terbukti efektif ketika dikombinasikan dengan strategi fine-tuning pada dataset berukuran terbatas [18].

DenseNet121 menerapkan konsep dense connection di mana setiap layer terhubung langsung ke semua layer berikutnya. Koneksi padat ini mendorong penggunaan ulang fitur, mengurangi jumlah parameter yang diperlukan, dan meningkatkan aliran gradien selama backpropagation. DenseNet121 memiliki 121 layer dengan sekitar 8 juta parameter, dan integrasinya dengan mekanisme atensi maupun optimasi hyperparameter telah terbukti meningkatkan performa pada klasifikasi gambar nyata dan AI [15], [16].

Pada setiap model, lapisan klasifikasi (top layer) dihapus dan digantikan dengan custom head yang terdiri dari: GlobalAveragePooling2D, BatchNormalization, Dense(256, ReLU) dengan L2 regularization ($\lambda=0.001$), Dropout(0.5), Dense(128, ReLU) dengan L2 regularization ($\lambda=0.001$), Dropout(0.3), dan Dense(1, Sigmoid) sebagai output untuk klasifikasi biner.

2.7 Strategi Pelatihan Dua Fase

Pelatihan model dilakukan dalam dua fase untuk mengoptimalkan transfer learning. Pada Fase 1 (Frozen Backbone), seluruh layer backbone pretrained dibekukan (`trainable=False`) dan hanya custom classification head yang dilatih. Optimizer yang digunakan adalah AdamW dengan learning rate 1×10^{-4} dan weight decay 1×10^{-5} . Loss function yang digunakan adalah Binary Crossentropy dengan label smoothing sebesar 0.1 untuk mencegah overconfidence model. Fase ini dijalankan selama maksimum 15 epoch.

Pada Fase 2 (Fine-tuning), 30 layer terakhir dari backbone diaktifkan kembali untuk fine-tuning (layer-layer awal tetap dibekukan). Learning rate diturunkan menjadi 1×10^{-5} dengan weight decay 1×10^{-6} , dan label smoothing dikurangi menjadi 0.05. Fine-tuning dijalankan selama maksimum 15 epoch tambahan. Callbacks yang digunakan pada kedua fase meliputi EarlyStopping (`patience=5`, `monitoring_val_loss`), ReduceLROnPlateau (`factor=0.2`, `patience=3`, `min_lr=1 \times 10^{-6}`), dan ModelCheckpoint untuk menyimpan bobot terbaik berdasarkan `val_accuracy`.

Class weight dihitung secara otomatis menggunakan `compute_class_weight` dengan parameter 'balanced' dari scikit-learn untuk menangani ketidakseimbangan kelas (AI=25,13%, Real=74,87%) pada data pelatihan. Bobot kelas yang lebih tinggi diberikan pada kelas minoritas (AI-Generated) agar model tidak bias terhadap kelas mayoritas. Mixed precision training (`mixed_float16`) diaktifkan untuk mempercepat komputasi pada GPU yang mendukung Tensor Cores.

Pemilihan nilai hyperparameter pada penelitian ini didasarkan pada praktik terbaik dari penelitian sebelumnya serta beberapa eksperimen awal. Untuk menjamin reproduktibilitas hasil, nilai *random seed* ditetapkan secara konsisten pada seluruh komponen yang bersifat stokastik, yaitu `random_state=42` pada proses pembagian data, serta `tf.random.set_seed(42)` dan `np.random.seed(42)` pada inisialisasi TensorFlow dan NumPy. Penetapan *seed* ini memastikan bahwa proses pelatihan dan evaluasi menghasilkan keluaran yang konsisten ketika eksperimen diulang.

2.8 Metrik Evaluasi

Evaluasi performa model dilakukan menggunakan lima metrik kuantitatif pada test set. Akurasi mengukur proporsi prediksi yang benar dari seluruh sampel. Presisi mengukur proporsi prediksi positif yang benar-benar positif. Recall (sensitivitas) mengukur proporsi sampel positif yang berhasil terdeteksi. F1-score merupakan rata-rata harmonik dari presisi dan recall, memberikan ukuran keseimbangan antara kedua metrik tersebut. AUC-ROC mengukur kemampuan diskriminatif model secara keseluruhan terlepas dari threshold keputusan. Penggunaan AUC-ROC sebagai metrik pendamping akurasi penting karena memberikan gambaran performa yang lebih andal pada kondisi data tidak seimbang [17]. Threshold keputusan yang digunakan adalah 0.5 untuk semua model.

2.9 Visualisasi Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) diimplementasikan untuk memberikan interpretabilitas visual terhadap keputusan model. Kebutuhan akan interpretabilitas ini menjadi penting dalam aplikasi forensik digital dan verifikasi media, karena sistem deteksi yang bersifat black-box sulit dipercaya tanpa penjelasan mengenai dasar pengambilan keputusannya [19]. Grad-CAM menghitung heatmap aktivasi dengan mengambil gradien dari skor kelas terhadap output layer konvolusi terakhir pada backbone. Heatmap yang dihasilkan dioverlay pada gambar asli menggunakan colormap JET dengan alpha blending 0.4 untuk menunjukkan area yang paling berpengaruh dalam pengambilan keputusan klasifikasi. Pendekatan yang menggabungkan ekstraksi fitur dengan visualisasi interpretabilitas semacam ini telah diterapkan pada berbagai domain klasifikasi citra untuk memverifikasi bahwa model berfokus pada fitur yang relevan [20].

3. HASIL DAN PEMBAHASAN

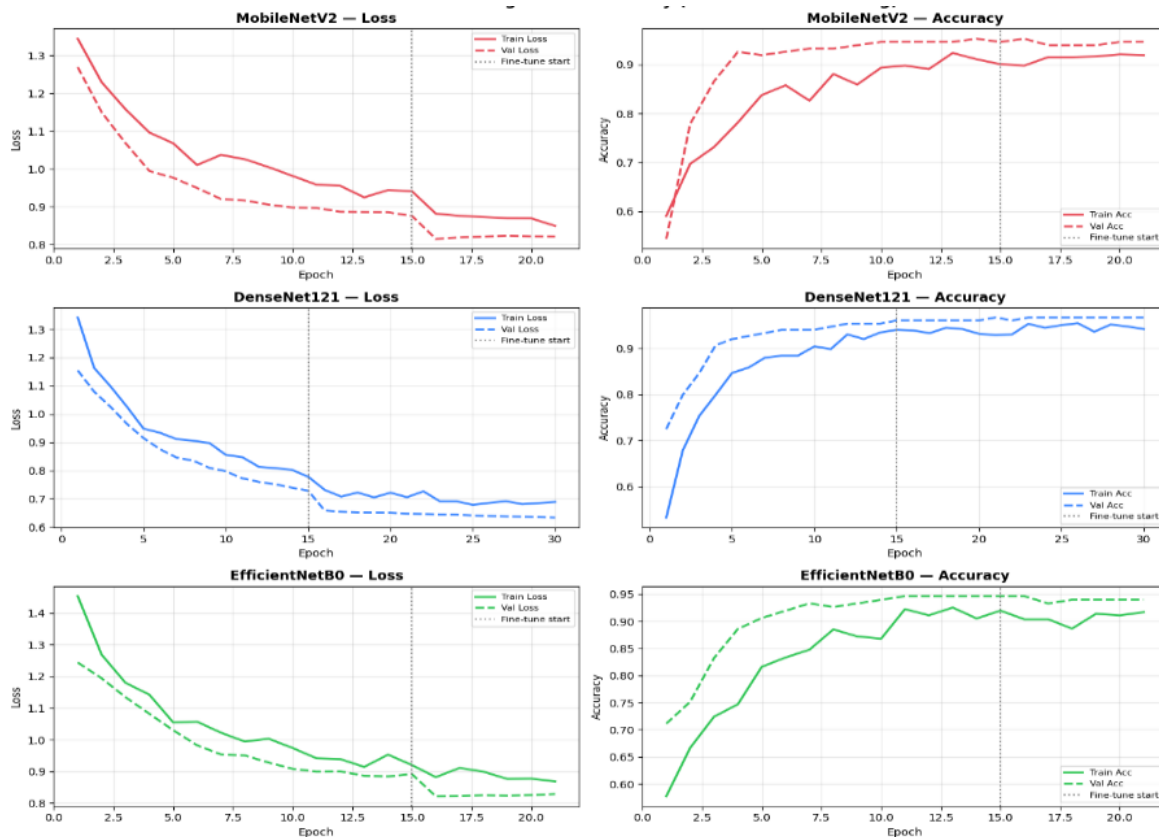
3.1 Lingkungan Eksperimen

Seluruh eksperimen dijalankan pada platform Kaggle Notebooks dengan akselerator GPU NVIDIA Tesla T4. Implementasi menggunakan TensorFlow 2.x dengan mixed precision training (`mixed_float16`) yang diaktifkan untuk mempercepat komputasi. Ukuran input citra ditetapkan 224×224 piksel dengan batch size 16. Ketiga model dilatih menggunakan strategi dua fase yang identik agar perbandingan bersifat adil, yaitu fase frozen backbone dilanjutkan dengan fase fine-tuning pada 30 layer terakhir. Komposisi data uji mengikuti pembagian pada subbab Dataset (Tabel 2).

3.2 Hasil Pelatihan Dua Fase

Proses pelatihan dipantau melalui kurva akurasi dan loss pada data pelatihan serta validasi di sepanjang epoch. Pada fase pertama (frozen backbone), ketiga model menunjukkan peningkatan akurasi validasi yang konsisten seiring bertambahnya epoch, dengan akurasi validasi mencapai kisaran 93–95% pada akhir fase. Pada fase kedua (fine-

tuning), pembukaan 30 layer terakhir backbone dengan learning rate yang lebih kecil (1×10^{-5}) memberikan peningkatan tambahan pada kemampuan diskriminatif model, sebagaimana terlihat dari nilai AUC validasi yang stabil di atas 0,98. Mekanisme EarlyStopping dan ReduceLROnPlateau bekerja efektif dalam mencegah overfitting, di mana pelatihan dihentikan secara otomatis ketika val_loss tidak lagi membaik dan bobot terbaik dipulihkan. Kurva pelatihan ketiga model ditunjukkan pada Gambar 3.



Gambar 3. Kurva akurasi dan loss pelatihan serta validasi ketiga model

3.3 Perbandingan Performa Model

Performa ketiga arsitektur dievaluasi menggunakan lima metrik pada data uji, yaitu akurasi, presisi, recall, F1-score, dan AUC-ROC. Hasil perbandingan lengkap disajikan pada Tabel 3. DenseNet121 mencapai performa tertinggi pada seluruh metrik, dengan akurasi 98,00%, presisi 99,10%, recall 98,21%, F1-score 98,65%, dan AUC-ROC 99,72%. EfficientNetB0 menempati posisi kedua dengan akurasi 95,33% dan AUC-ROC 99,15%, sedangkan MobileNetV2 memperoleh akurasi 94,67% dan AUC-ROC 98,97%.

Tabel 3. Perbandingan Performa Tiga Model pada Data Uji

Model	Akurasi (%)	Presisi (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)	Waktu Latih (s)	Parameter (juta)
DenseNet121	98,00	99,10	98,21	98,65	99,72	806,88	7,34
EfficientNetB0	95,33	99,07	94,64	96,80	99,15	688,78	4,42
MobileNetV2	94,67	99,06	93,75	96,33	98,97	623,26	2,62

Dari sisi efisiensi komputasi, MobileNetV2 mencatat waktu pelatihan tercepat (623,26 detik), diikuti EfficientNetB0 (688,78 detik) dan DenseNet121 (806,88 detik). DenseNet121 memerlukan waktu sekitar 29% lebih lama dibandingkan MobileNetV2 sebagai konsekuensi dari kompleksitas arsitektur yang lebih tinggi akibat mekanisme dense connectivity yang menghubungkan setiap layer dengan seluruh layer sebelumnya. Perbedaan ini menjadi pertimbangan penting dalam pemilihan model, terutama untuk skenario dengan keterbatasan sumber daya atau kebutuhan inferensi cepat.

Keunggulan DenseNet121 dapat dikaitkan dengan mekanisme dense connectivity yang mendorong penggunaan ulang fitur dan memperkuat aliran gradien, sehingga model mampu menangkap fitur diskriminatif yang lebih kaya untuk membedakan citra AI dari citra nyata. Namun demikian, keunggulan ini disertai dengan biaya komputasi yang lebih tinggi, terlihat dari jumlah parameter (7,34 juta) dan waktu pelatihan (806,88 detik) yang paling besar di antara ketiga model. Temuan ini konsisten dengan penelitian Hasaniyyah et al. [16] yang menunjukkan bahwa

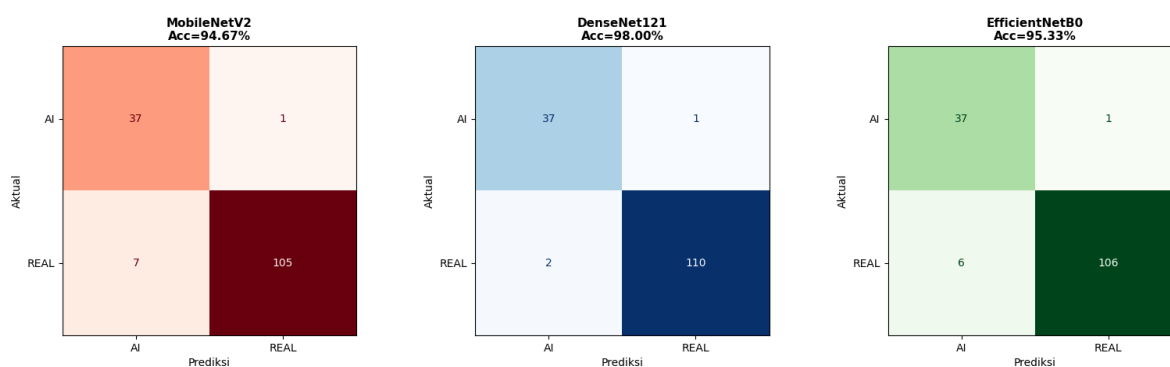
DenseNet121 mampu mempertahankan fitur diskriminatif lebih baik dibandingkan arsitektur ringan seperti MobileNet, sehingga menghasilkan performa klasifikasi yang lebih tinggi pada tugas pembedaan citra nyata dan citra AI.

Sebaliknya, MobileNetV2 sebagai model paling ringan (2,62 juta parameter) dengan waktu pelatihan tercepat (623,26 detik) tetap mempertahankan performa yang kompetitif dengan akurasi 94,67% dan AUC-ROC 98,97%. Hal ini menunjukkan bahwa untuk skenario deployment dengan keterbatasan sumber daya komputasi, MobileNetV2 menawarkan trade-off yang menarik antara efisiensi dan akurasi. EfficientNetB0 menempati posisi tengah baik dari segi performa maupun kompleksitas, menjadikannya alternatif seimbang.

Satu pola menarik yang konsisten pada ketiga model adalah nilai presisi yang sangat tinggi (di atas 99%) yang melampaui nilai recall. Hal ini mengindikasikan bahwa ketika model memprediksi suatu citra sebagai Real, prediksi tersebut hampir selalu benar, sementara terdapat sebagian kecil citra Real yang keliru diklasifikasikan sebagai AI. Pola ini relevan dalam konteks deteksi gambar AI, di mana meminimalkan kesalahan klasifikasi menjadi prioritas penting.

3.4 Analisis Confusion Matrix

Confusion matrix memberikan gambaran rinci mengenai distribusi kesalahan klasifikasi setiap model, sebagaimana ditunjukkan pada Gambar 4. Pada DenseNet121, dari 38 citra AI seluruhnya terklasifikasi dengan baik di mana 37 citra dikenali benar sebagai AI, sementara dari 112 citra Real sebanyak 110 dikenali benar dan hanya 2 yang keliru diklasifikasikan sebagai AI. EfficientNetB0 mengklasifikasikan 37 dari 38 citra AI dengan benar dan 106 dari 112 citra Real dengan benar. MobileNetV2 menunjukkan pola serupa dengan 37 dari 38 citra AI dan 105 dari 112 citra Real terklasifikasi dengan benar.

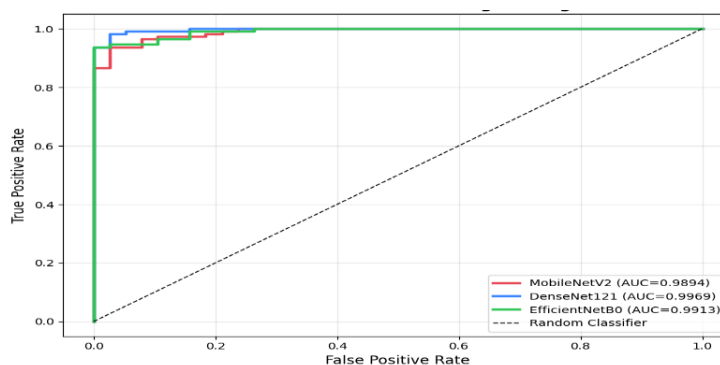


Gambar 4. Confusion matrix ketiga model pada data uji

Analisis confusion matrix mengonfirmasi bahwa kesalahan klasifikasi terbesar terjadi pada citra Real yang diprediksi sebagai AI (false negative terhadap kelas Real), bukan sebaliknya. DenseNet121 menunjukkan jumlah kesalahan paling sedikit dengan total hanya 3 kesalahan dari 150 citra uji, konsisten dengan akurasi yang tertinggi.

3.5 Analisis Kurva ROC

Kurva Receiver Operating Characteristic (ROC) digunakan untuk mengevaluasi kemampuan diskriminatif model pada berbagai nilai threshold. Sebagaimana ditunjukkan pada Gambar 5, ketiga model menghasilkan kurva ROC yang mendekati sudut kiri atas, menandakan kemampuan klasifikasi yang sangat baik. DenseNet121 memperoleh nilai AUC tertinggi sebesar 0,9972, diikuti EfficientNetB0 dengan 0,9915 dan MobileNetV2 dengan 0,9897. Nilai AUC yang konsisten di atas 0,98 untuk ketiga model menunjukkan bahwa seluruh arsitektur memiliki kemampuan pemisahan kelas yang andal terlepas dari pemilihan threshold, yang menjadi indikator penting keandalan model pada kondisi data yang tidak seimbang.



Gambar 5. Kurva ROC ketiga model

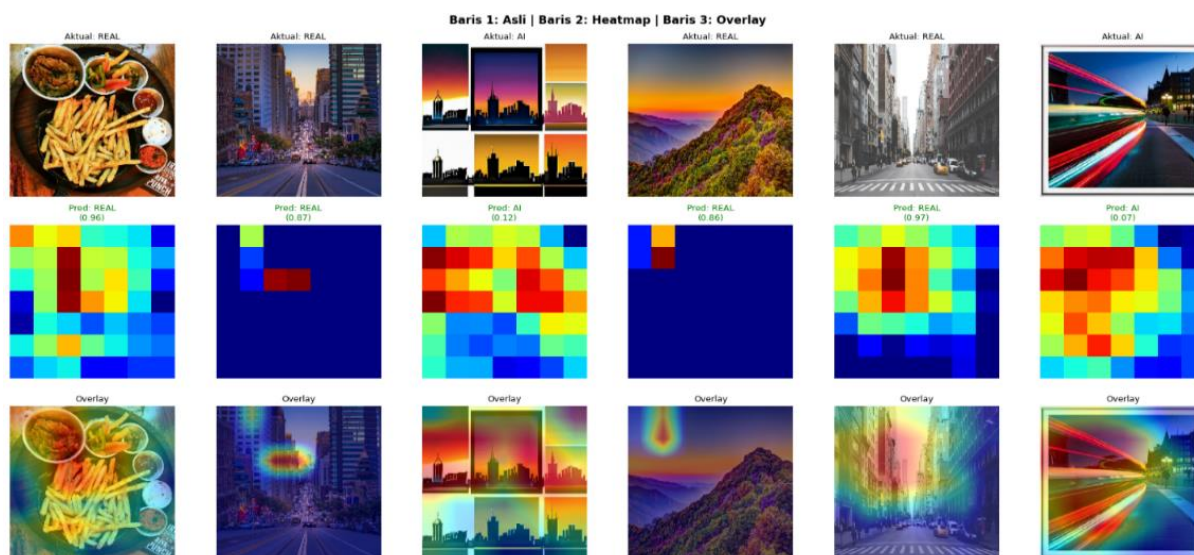
3.6 Distribusi Confidence Score

Distribusi skor confidence prediksi dianalisis untuk memahami tingkat keyakinan model dalam pengambilan keputusan. Ketiga model menunjukkan pemisahan yang jelas antara distribusi skor untuk citra AI dan citra Real, di mana sebagian besar prediksi terkonsentrasi pada nilai ekstrem (mendekati 0 untuk AI dan mendekati 1 untuk Real). Pola ini mengindikasikan bahwa model tidak hanya menghasilkan klasifikasi yang benar, tetapi juga melakukannya dengan tingkat keyakinan yang tinggi. Penggunaan label smoothing selama pelatihan turut berkontribusi mencegah model menjadi terlalu yakin (overconfident) secara berlebihan, sehingga skor confidence yang dihasilkan tetap terkalibrasi dengan baik.

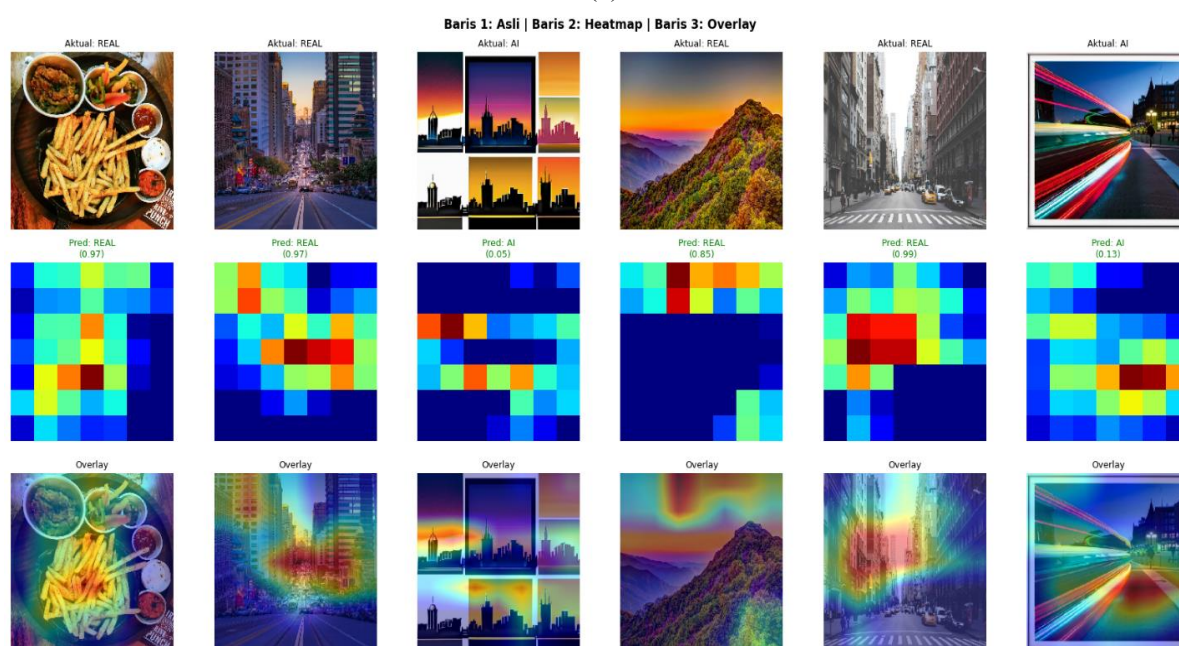
3.7 Analisis Interpretabilitas Grad-CAM

Untuk meningkatkan transparansi dan interpretabilitas model, visualisasi Grad-CAM diterapkan guna mengidentifikasi wilayah citra yang menjadi dasar keputusan klasifikasi. Gambar 7 menampilkan hasil Grad-CAM pada ketiga model, yang masing-masing terdiri dari tiga baris: citra asli, heatmap aktivasi, dan overlay heatmap pada citra asli. Heatmap divisualisasikan menggunakan colormap JET, di mana wilayah berwarna merah menunjukkan tingkat kontribusi tertinggi terhadap keputusan model, sedangkan wilayah biru menunjukkan kontribusi terendah.

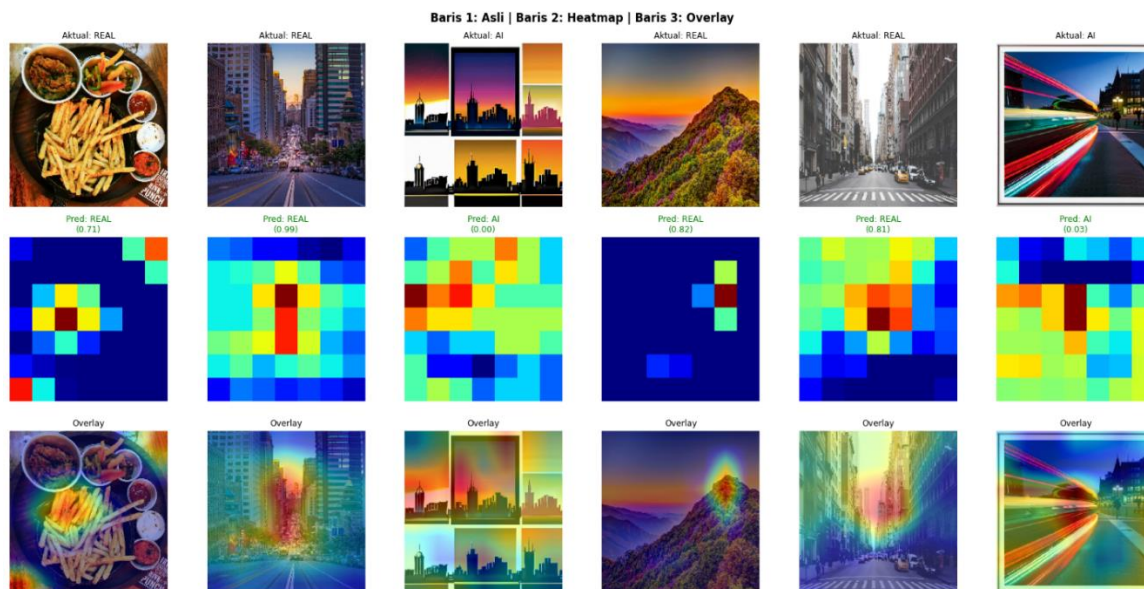
(a)



(b)



(c)

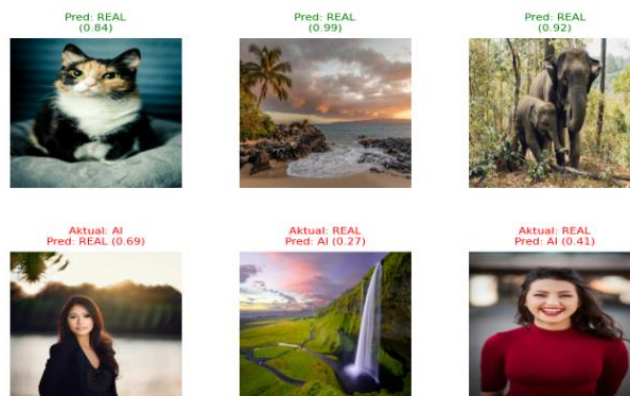


Gambar 7. Visualisasi Grad-CAM ketiga model (a) MobileNetV2, (b) DenseNet121, (c) EfficientNetB0

Hasil visualisasi menunjukkan bahwa ketiga model memusatkan perhatian pada wilayah yang koheren dan relevan secara semantik, bukan menyebar secara acak di seluruh citra. Pada citra nyata yang memuat objek utama yang jelas, seperti makanan, bangunan, maupun kendaraan, area aktivasi terkonsentrasi pada objek utama tersebut, menandakan bahwa model berdasarkan keputusannya pada konten citra dan bukan pada latar belakang semata. Sementara itu, pada citra yang dihasilkan AI, area aktivasi cenderung terfokus pada wilayah dengan transisi warna gradien dan tekstur sintetis, yang merupakan lokasi munculnya artefak khas proses generasi. Pola ini sejalan dengan temuan literatur bahwa artefak halus pada tekstur dan transisi warna menjadi penanda penting yang membedakan citra sintetis dari citra nyata.

Perbandingan antar model memperlihatkan bahwa DenseNet121, sebagai model dengan performa tertinggi, menghasilkan peta aktivasi yang lebih terfokus dan tajam pada wilayah diskriminatif dibandingkan kedua model lainnya, yang konsisten dengan akurasi yang superior. Temuan ini memberikan bukti kualitatif bahwa keputusan model tidak bersifat sembarang, melainkan didasarkan pada fitur visual yang dapat ditelusuri secara spasial, sehingga memperkuat kepercayaan terhadap keandalan model dalam aplikasi forensik digital dan verifikasi media.

Selain analisis Grad-CAM, contoh prediksi benar dan salah pada model terbaik disajikan pada Gambar 8 untuk memberikan gambaran kualitatif mengenai perilaku model. Baris pertama menampilkan kasus prediksi benar dengan tingkat keyakinan yang tinggi, sedangkan baris kedua menampilkan kasus kesalahan klasifikasi. Analisis terhadap kasus kesalahan menunjukkan bahwa kekeliruan umumnya terjadi pada citra AI berkualitas sangat tinggi, khususnya gambar wajah manusia dan pemandangan yang sangat realistis, yang keliru diklasifikasikan sebagai citra nyata. Hal ini mengindikasikan bahwa tantangan utama model terletak pada citra sintetis modern yang memiliki tingkat fotorealisme tinggi sehingga jejak artefak generatifnya menjadi sangat minim.



Gambar 8. Contoh prediksi benar dan salah pada model terbaik (DenseNet121)

3.8 Uji Signifikansi Statistik

Untuk menguji apakah perbedaan performa antar model bersifat signifikan secara statistik, dilakukan McNemar test pada pasangan prediksi ketiga model di data uji. Hasil pengujian disajikan pada Tabel 4.

Tabel 4. Hasil McNemar Test Antar Model

Perbandingan	b	c	p-value	Signifikan ($\alpha=0,05$)
MobileNetV2 vs DenseNet121	1	6	0,1250	Tidak
MobileNetV2 vs EfficientNetB0	2	3	1,0000	Tidak
DenseNet121 vs EfficientNetB0	5	1	0,2188	Tidak

Hasil McNemar test menunjukkan bahwa seluruh pasangan perbandingan memiliki nilai p-value di atas 0,05, sehingga perbedaan performa antar ketiga model tidak signifikan secara statistik. Hal ini mengindikasikan bahwa meskipun DenseNet121 mencatat akurasi tertinggi secara numerik, ketiga arsitektur memiliki kemampuan klasifikasi yang setara secara statistik pada dataset ini. Temuan ini sebagian disebabkan oleh ukuran data uji yang relatif kecil (150 citra), sehingga jumlah kasus prediksi yang berbeda antar model terlalu sedikit untuk mencapai signifikansi statistik. Implikasinya, pemilihan model dalam praktik dapat mempertimbangkan faktor efisiensi komputasi, di mana MobileNetV2 dengan jumlah parameter paling sedikit menjadi pilihan menarik tanpa penurunan performa yang berarti.

3.9 Validasi Silang (Cross Validation)

Untuk menguji kestabilan dan keandalan model terbaik, dilakukan 5-fold stratified cross validation pada DenseNet121 menggunakan keseluruhan dataset. Hasil validasi silang disajikan pada Tabel 5.

Tabel 5. Hasil 5-Fold Cross Validation DenseNet121

Fold	Accuracy (%)	F1-Score (%)	AUC-ROC (%)
1	96,48	97,61	99,79
2	95,48	96,99	99,06
3	96,98	97,95	99,89
4	96,98	97,99	99,44
5	98,49	99,00	99,57
Rata-rata	96,88 $\pm$ 0,97	97,91 $\pm$ 0,65	99,55 $\pm$ 0,29

DenseNet121 memperoleh akurasi rata-rata 96,88% dengan standar deviasi 0,97%, F1-score 97,91% ($\pm 0,65\%$), dan AUC-ROC 99,55% ($\pm 0,29\%$). Nilai standar deviasi yang kecil pada seluruh metrik menunjukkan bahwa performa model konsisten dan stabil di seluruh fold, sehingga hasil yang diperoleh tidak bergantung pada pembagian data tertentu. Konsistensi ini memperkuat keandalan DenseNet121 sebagai model deteksi citra AI dan mengonfirmasi bahwa performa tinggi yang dicapai bukan merupakan akibat dari pembagian data yang menguntungkan secara kebetulan. Validasi silang difokuskan pada DenseNet121 sebagai model dengan performa terbaik guna menguji kestabilan dan keandalannya secara mendalam. Pembatasan ini didasarkan pada pertimbangan keterbatasan sumber daya komputasi dan waktu pelatihan pada lingkungan Kaggle Notebooks, mengingat setiap model memerlukan pelatihan dua fase yang diulang sebanyak lima kali (lima fold). Pengujian cross validation terhadap EfficientNetB0 dan MobileNetV2 direkomendasikan sebagai bagian dari penelitian selanjutnya.

3.10 Ablation Study

Untuk menganalisis kontribusi masing-masing komponen terhadap performa model, dilakukan ablation study pada DenseNet121 dengan menambahkan komponen secara kumulatif. Hasil disajikan pada Tabel 6.

Tabel 6. Hasil Ablation Study DenseNet121

Konfigurasi	Accuracy (%)
Baseline (tanpa augmentasi)	96,67
+ Augmentasi	95,33
+ Class Weight	94,00
+ Label Smoothing	95,33
+ Fine-Tuning	94,67

Hasil ablation study menunjukkan bahwa konfigurasi baseline tanpa augmentasi memperoleh akurasi tertinggi (96,67%) pada satu kali pembagian data uji, sementara penambahan komponen augmentasi, class weight, dan label smoothing tidak menaikkan akurasi pada data uji tunggal ini. Walaupun akurasi pada satu kali pembagian data sedikit menurun, seluruh konfigurasi memberikan performa yang lebih stabil berdasarkan hasil cross validation. Hal ini karena komponen augmentasi, class weight, dan label smoothing berperan utama sebagai mekanisme regularisasi yang manfaatnya lebih terlihat pada kemampuan generalisasi dan kestabilan model dibandingkan pada akurasi mentah di satu kali pembagian data. Fenomena ini juga dipengaruhi oleh ukuran data uji yang kecil (150 citra), sehingga selisih beberapa citra saja sudah berdampak besar pada nilai akurasi. Dengan demikian, meskipun secara numerik komponen-komponen tersebut tidak menaikkan akurasi pada data uji tunggal, perannya tetap penting dalam menjaga ketahanan dan keterkalibrasian model, terutama pada kondisi data tidak seimbang.

3.11 Pembahasan

Untuk menempatkan hasil penelitian ini dalam konteks yang lebih luas, dilakukan perbandingan dengan beberapa penelitian sebelumnya pada tugas deteksi citra AI. Untuk menempatkan hasil penelitian ini dalam konteks yang lebih luas, dilakukan perbandingan dengan beberapa penelitian sebelumnya pada tugas deteksi citra AI.

Perbandingan dengan penelitian sebelumnya menunjukkan bahwa DenseNet121 pada penelitian ini, yang dilatih pada dataset AI vs Real Images, mencapai akurasi 98,00%. Nilai ini lebih tinggi dibandingkan pendekatan CNN dasar oleh Bird & Lotfi [14] yang memperoleh akurasi 92,98% pada dataset CIFAKE, serta berada pada rentang yang setara dengan Hasaniyyah et al. [16] yang mencapai 98,52% menggunakan DenseNet-121 dengan penambahan mekanisme Squeeze-and-Excitation pada dataset CIFAKE. Perlu dicatat bahwa perbandingan ini bersifat indikatif karena setiap penelitian menggunakan dataset dan protokol evaluasi yang berbeda. Meskipun demikian, hasil ini menunjukkan bahwa pendekatan transfer learning dua fase pada DenseNet121 tanpa modifikasi arsitektur tambahan mampu menghasilkan performa yang kompetitif dengan metode yang menggunakan mekanisme atensi tambahan, sekaligus menegaskan bahwa strategi pelatihan yang tepat dapat menjadi faktor yang sama pentingnya dengan kompleksitas arsitektur.

4. KESIMPULAN

Penelitian ini berhasil membandingkan performa tiga arsitektur CNN *pretrained*, yaitu DenseNet121, EfficientNetB0, dan MobileNetV2, dalam mendeteksi *AI-generated images* menggunakan pendekatan *transfer learning* dua fase, di mana DenseNet121 memperoleh performa tertinggi dengan akurasi 98,00%, presisi 99,10%, *recall* 98,21%, F1-score 98,65%, dan AUC-ROC 99,72%, serta terbukti stabil melalui *5-fold cross validation* dengan akurasi rata-rata 96,88% ($\pm 0,97\%$), sementara EfficientNetB0 dan MobileNetV2 menghasilkan performa kompetitif dengan akurasi masing-masing 95,33% dan 94,67%. Uji McNemar menunjukkan perbedaan performa antar ketiga model tidak signifikan secara statistik, sehingga ketiganya dapat dianggap setara dan pemilihan model dapat mempertimbangkan efisiensi komputasi, dengan MobileNetV2 unggul karena jumlah parameter paling sedikit, sedangkan visualisasi Grad-CAM mengonfirmasi interpretabilitas model yang memusatkan perhatian pada objek utama citra serta wilayah bertekstur dan bertransisi warna yang menjadi penanda artefak generatif. Meskipun demikian, penelitian ini memiliki keterbatasan berupa ukuran dataset yang relatif kecil (995 citra) dibanding *benchmark* berskala besar dan data uji yang terbatas sehingga uji signifikansi statistik kurang konklusif, serta kesalahan klasifikasi yang umumnya terjadi pada citra AI berfotorealisme tinggi. Oleh karena itu, penelitian selanjutnya disarankan menguji generalisasi model pada dataset yang lebih besar dan beragam, mengeksplorasi arsitektur yang lebih mutakhir seperti Vision Transformer, serta mengevaluasi ketahanan model terhadap berbagai *postprocessing* seperti kompresi JPEG dan penambahan *noise* guna memastikan kesiapannya dalam penerapan di dunia nyata.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: N. A. H., dan S.;

Metodologi: N. A. H., dan S.;

Perangkat Lunak: N. A. H., dan S.;

Validasi: N. A. H., dan S.;

Analisis Formal: N. A. H., dan S.;

Investigasi: N. A. H., dan S.;

Sumber Daya: N. A. H., dan S.;

Kurasi Data: N. A. H., dan S.;

Penulisan Draf Awal: N. A. H.;

Penulisan, Peninjauan, dan Penyuntingan: N. A. H., dan S.;

Visualisasi: N. A. H.;

Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis korespondensi.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan eksternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

- [1] L. Y. Gong and X. J. Li, "A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges," *Electronics (Basel)*, vol. 13, no. 3, p. 585, Jan. 2024, doi: 10.3390/electronics13030585.
- [2] Z. Akhtar, "Deepfakes Generation and Detection: A Short Survey," *J. Imaging*, vol. 9, no. 1, p. 18, Jan. 2023, doi: 10.3390/jimaging9010018.
- [3] A. López-Borrull and C. Lopezosa, "Mapping the Impact of Generative AI on Disinformation: Insights from a Scoping Review," *Publications*, vol. 13, no. 3, p. 33, Jul. 2025, doi: 10.3390/publications13030033.
- [4] Á. F. Gambín, A. Yazidi, A. Vasilakos, H. Haugerud, and Y. Djenouri, "Deepfakes: current and future trends," *Artif. Intell. Rev.*, vol. 57, no. 3, p. 64, Feb. 2024, doi: 10.1007/s10462-023-10679-x.
- [5] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 2, Mar. 2024, doi: 10.1002/widm.1520.
- [6] Q. Bammey, "Synthbuster: Towards Detection of Diffusion Model Generated Images," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 1–9, 2024, doi: 10.1109/OJSP.2023.3337714.
- [7] R. Mubarak, T. Alsaboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," *IEEE Access*, vol. 11, pp. 144497–144529, 2023, doi: 10.1109/ACCESS.2023.3344653.
- [8] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–41, Jan. 2022, doi: 10.1145/3425780.
- [9] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A Survey on Deepfake Video Detection," *IET Biom.*, vol. 10, no. 6, pp. 607–624, Nov. 2021, doi: 10.1049/bme2.12031.
- [10] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, Feb. 2023, doi: 10.1007/s10489-022-03766-z.
- [11] R. Yumlembam, B. Issac, N. Aslam, E. K. Babu, J. Collyer, and F. Kennedy, "Detection of AI generated images using combined uncertainty measures and particle swarm optimised rejection mechanism," *Sci. Rep.*, vol. 15, no. 1, p. 44021, Nov. 2025, doi: 10.1038/s41598-025-28572-8.
- [12] Z. Meng, B. Peng, J. Dong, T. Tan, and H. Cheng, "Artifact feature purification for cross-domain detection of AI-generated images," *Computer Vision and Image Understanding*, vol. 247, p. 104078, Oct. 2024, doi: 10.1016/j.cviu.2024.104078.
- [13] A. M. Almars, "Deepfakes Detection Techniques Using Deep Learning: A Survey," *Journal of Computer and Communications*, vol. 09, no. 05, pp. 20–35, 2021, doi: 10.4236/jcc.2021.95003.
- [14] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," *IEEE Access*, vol. 12, pp. 15642–15650, 2024, doi: 10.1109/ACCESS.2024.3356122.
- [15] A. R. Gunukula, H. Das Gupta, and V. S. Sheng, "Detecting AI-Generated Images Using a Hybrid ResNet-SE Attention Model," *Applied Sciences*, vol. 15, no. 13, p. 7421, Jul. 2025, doi: 10.3390/app15137421.
- [16] N. Hasaniyyah, K. Khadijah, S. Sutikno, and Z. A. Tyas, "Integration of Squeeze-and-Excitation in Densenet-121 for Classifying Real and AI-Generated Images," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 6, pp. 5474–5488, Dec. 2025, doi: 10.52436/1.jutif.2025.6.6.5336.
- [17] G. Chen, C. Du, Y. Yu, H. Hu, H. Duan, and H. Zhu, "A Deepfake Image Detection Method Based on a Multi-Graph Attention Network," *Electronics (Basel)*, vol. 14, no. 3, p. 482, Jan. 2025, doi: 10.3390/electronics14030482.
- [18] A. N. Abiyyu and M. Rahardi, "Comparison of Transfer learning Models MobileNetV3-Large and EfficientNet-B0 for Rice Leaf Disease Classification," vol. 10, no. 1, pp. 818–828, 2026.
- [19] I. Amerini *et al.*, "Deepfake Media Forensics: Status and Future Challenges," *J. Imaging*, vol. 11, no. 3, p. 73, Feb. 2025, doi: 10.3390/jimaging11030073.
- [20] K. A. Huang, A. Alkadri, and N. Prakash, "Employing Squeeze-and-Excitation Architecture in a Fine-Tuned Convolutional Neural Network for Magnetic Resonance Imaging Tumor Classification," *Cureus*, Mar. 2025, doi: 10.7759/cureus.80084.