

Deteksi Anomali Tekstural pada Gambar Generatif AI menggunakan Convolutional Neural Network (CNN) dengan Arsitektur EfficientNetB2

Farhan Muhammad^{*1}, Marcelia Anggiriani², Ahmad Yusuf Roky³, Tri Tia Fahada⁴, Andre Pratama⁵

^{1,2,3,4,5}Informatika, Fakultas Teknologi dan Ilmu Komputer, Universitas Satya Terra Bhinneka, Indonesia

Email: ¹farhan.muhammd19@gmail.com, ²maranggi14@gmail.com, ³ahmadyusufroky85@gmail.com, ⁴tritiahahada123@gmail.com, ⁵andrepratama@satyaterrabhinneka.ac.id

Abstrak

Perkembangan teknologi *image generation* berbasis *Artificial Intelligence* (AI) membuat batas antara gambar buatan AI dan gambar *Non-AI* semakin sulit dibedakan. Kondisi ini mendorong kebutuhan akan sistem deteksi otomatis yang akurat. Penelitian ini bertujuan mengembangkan model klasifikasi citra untuk membedakan gambar AI dan *Non-AI* menggunakan arsitektur *EfficientNet-B2*. Evaluasi model dilakukan menggunakan dataset CIFAKE yang terdiri dari citra REAL dan FAKE sebagai representasi gambar non-AI dan AI. Model dibangun dengan memanfaatkan bobot *pretrained* sebagai *feature extractor* dan menambahkan lapisan klasifikasi dua kelas. Pelatihan dilakukan dalam dua tahap menggunakan *Adam optimizer* untuk meningkatkan stabilitas dan kemampuan generalisasi. Hasil penelitian menunjukkan bahwa *EfficientNet-B2* mampu mengekstraksi fitur visual secara efektif. Pada pelatihan tahap awal, model memperoleh akurasi validasi 85,37%, kemudian meningkat menjadi 86,11% setelah optimisasi, dan mencapai performa terbaik pada tahap *fine-tuning* dengan akurasi 86,67%. Evaluasi pada data pengujian menghasilkan akurasi 87,12% dengan nilai *precision*, *recall*, dan *f1-score* yang seimbang pada kedua kelas. Secara keseluruhan, model yang dibangun menunjukkan performa yang stabil dan akurat dalam mendeteksi gambar AI. Studi ini membuktikan bahwa pendekatan *transfer learning* dan *fine-tuning* dengan *EfficientNet-B2* merupakan metode yang efektif untuk tugas klasifikasi gambar dua kelas.

Kata kunci: *AI-generated images, deep learning, EfficientNet-B2, klasifikasi gambar, transfer learning*

Abstract

The rapid development of AI-based image generation technologies has increasingly blurred the distinction between AI-generated images and real photographs. This creates a strong need for an automatic and reliable detection system. This study aims to develop an image classification model capable of distinguishing AI-generated and non-AI images using the EfficientNet-B2 architecture. The model is evaluated using the CIFAKE dataset, which consists of REAL and FAKE image classes representing non-AI and AI-generated images. The model utilizes pretrained weights as the main feature extractor, with additional classification layers for two output classes. Training is conducted in two stages using the Adam optimizer to improve stability and generalization performance. The results show that EfficientNet-B2 effectively extracts visual features for classification. In the baseline training stage, the model achieves a validation accuracy of 85.37%, which improves to 86.11% after optimization, and reaches its best performance during fine-tuning with a validation accuracy of 86.67%. Evaluation on the test dataset yields an accuracy of 87.12%, with balanced precision, recall, and f1-score across both classes. Overall, the proposed model demonstrates stable and accurate performance in detecting AI-generated images. The study confirms that transfer learning combined with fine-tuning of EfficientNet-B2 is an effective approach for binary image classification tasks.

Keywords: *AI-generated images, deep learning, EfficientNet-B2, image classification, transfer learning*

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)



1. PENDAHULUAN

Perkembangan kecerdasan buatan (AI), khususnya di bidang generatif, telah mengalami percepatan signifikan dalam beberapa tahun terakhir dan memberikan dampak besar pada berbagai sektor, termasuk industri kreatif. Munculnya teknologi generative image AI seperti DALL·E 2 dan Stable Diffusion memungkinkan pembuatan gambar berkualitas tinggi hanya dengan perintah teks, sehingga merevolusi

dan mempercepat proses produksi konten visual[1]. Aplikasinya telah membawa manfaat nyata dalam bidang desain, ilustrasi, dan manipulasi wajah (Face Attribute Manipulation)[2]. Namun, di balik kemudahan dan efisiensi yang ditawarkan, kemajuan ini juga memunculkan tantangan baru yang kompleks terkait dengan kualitas, orisinalitas karya, dan potensi penyalahgunaan. Teknologi deepfake, sebagai contoh, dapat dimanfaatkan untuk membuat konten palsu yang berpotensi menimbulkan penipuan, manipulasi politik, dan penyebaran misinformasi[3]. Di sisi lain, meningkatnya realisme gambar generatif membuat masyarakat semakin sulit membedakan gambar asli (real) dan gambar sintetis (synthetic).

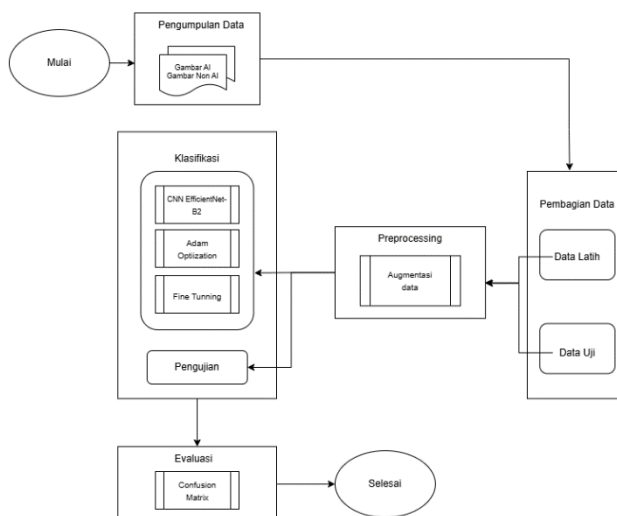
Studi terkini menunjukkan bahwa kemampuan manusia rata-rata dalam mengidentifikasi gambar generatif hanya sekitar 61%, jauh dari prediksi 85%[4]. Fakta ini menegaskan kebutuhan mendesak akan teknologi deteksi otomatis yang lebih akurat dan andal. Salah satu pendekatan yang menjanjikan adalah deteksi berbasis anomali tekstural, karena gambar yang dihasilkan AI sering kali menunjukkan ketidakselarasan tekstur akibat tidak merepresentasikan interaksi fisik yang alami antara pencahayaan, material, dan karakteristik kamera. Teknik analisis tekstur seperti Gray Level Co-occurrence Matrix (GLCM) telah terbukti efektif dalam mengidentifikasi pola keabnormalan tersebut[5]. Untuk mendukung deteksi semacam itu, metode Convolutional Neural Network (CNN) dinilai sangat relevan karena kemampuannya yang unggul dalam mengekstraksi pola visual dan tekstural secara mendalam. EfficientNetB2, dengan pendekatan compound scaling yang menyeimbangkan kedalaman, lebar, dan resolusi jaringan, menawarkan keunggulan untuk mempelajari detail tekstural secara efisien tanpa membutuhkan sumber daya komputasi yang berlebihan dibandingkan arsitektur yang lebih kompleks.

Penelitian-penelitian terdahulu mengonfirmasi keefektifan CNN, khususnya keluarga EfficientNet, dalam tugas klasifikasi berbasis tekstur. Studi oleh [6] menunjukkan akurasi tinggi (96,78%) pada klasifikasi citra wayang kulit, sementara penelitian lain oleh [7] mencapai akurasi 96% dalam membedakan motif batik. Aplikasi di domain lain, seperti deteksi penyakit daun oleh [8], juga menghasilkan akurasi hingga 98% menggunakan EfficientNetB3. Temuan-konsisten ini menguatkan landasan bahwa arsitektur EfficientNet memiliki potensi besar dan relevan untuk diterapkan dalam mendeteksi anomali tekstural pada gambar generatif AI yang menjadi fokus penelitian ini.

Adapun tujuan dari penelitian ini adalah untuk mengembangkan model CNN berbasis EfficientNetB2 untuk melakukan deteksi anomali tekstural pada gambar generatif AI serta menganalisis dan mengevaluasi performa model dalam mengenali perbedaan tekstural antara gambar asli dan gambar generatif menggunakan metrik evaluasi seperti akurasi, precision, recall, dan f1-score.

2. METODE PENELITIAN

Penelitian ini terdiri dari enam tahapan utama yang dirancang secara sistematis untuk membangun model deteksi anomali tekstural pada gambar generatif AI berbasis EfficientNetB2 secara akurat dan efisien. Alur tahapan tersebut dapat dilihat pada **Gambar 1**.



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

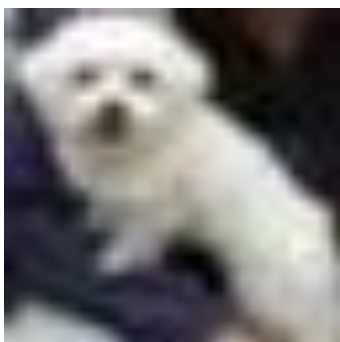
Dataset yang digunakan dalam penelitian ini adalah **CIFAKE: Real and AI-Generated Synthetic Images**, sebuah dataset *open-source* yang dipublikasikan oleh Bird dan Lotfi [9] dan tersedia melalui platform Kaggle. Dataset CIFAKE dirancang secara khusus untuk mendukung tugas klasifikasi citra antara gambar nyata (REAL) dan gambar sintesis hasil generasi kecerdasan buatan (FAKE).

Dataset CIFAKE terdiri atas dua kelas utama, yaitu **REAL dan FAKE**. Citra REAL bersumber dari dataset CIFAR-10 yang diperkenalkan oleh Krizhevsky dan Hinton [10], sedangkan citra FAKE merupakan gambar sintesis yang dihasilkan menggunakan model generatif Stable Diffusion versi 1.4, dengan karakteristik visual yang disesuaikan agar menyerupai distribusi citra CIFAR-10.

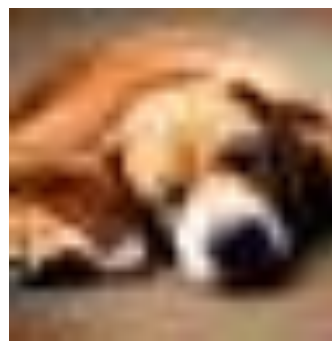
Secara keseluruhan, dataset CIFAKE menyediakan 120.000 citra, yang terbagi menjadi 100.000 citra untuk data pelatihan (masing-masing 50.000 citra per kelas) dan 20.000 citra untuk data pengujian (masing-masing 10.000 citra per kelas). Seluruh citra memiliki format warna RGB dengan resolusi asli 32×32 piksel, sesuai dengan spesifikasi standar dataset CIFAR-10.

Penggunaan dataset CIFAKE memungkinkan penelitian ini dilakukan secara terkontrol karena dataset tersebut telah memiliki *ground truth* yang jelas, distribusi kelas yang seimbang, serta telah digunakan dan divalidasi dalam penelitian sebelumnya terkait identifikasi citra sintesis berbasis *deep learning* [9]. Meskipun sumber citra REAL berasal dari dataset CIFAR-10 yang diperkenalkan pada tahun 2009 [10], sitasi tersebut dicantumkan untuk menjelaskan asal dan karakteristik data yang digunakan. Fokus metodologi dan kontribusi ilmiah penelitian ini tetap didasarkan pada penggunaan dan analisis dataset CIFAKE serta pendekatan deteksi citra sintesis yang relevan dengan perkembangan penelitian terkini.

Untuk memberikan gambaran visual mengenai karakteristik dataset yang digunakan, contoh citra REAL yang bersumber dari CIFAR-10 dan citra FAKE hasil generasi Stable Diffusion v1.4 masing-masing ditampilkan pada **Gambar 2** dan **Gambar 3**. Visualisasi ini menunjukkan perbedaan karakteristik visual antara citra nyata dan citra sintesis pada dataset CIFAKE.



Gambar 2. Contoh dataset CIFAKE citra REAL dari dataset CIFAR-10.



Gambar 3. Contoh dataset CIFAKE citra FAKE hasil generasi Stable Diffusion v1.4

2.2 Pembagian Data

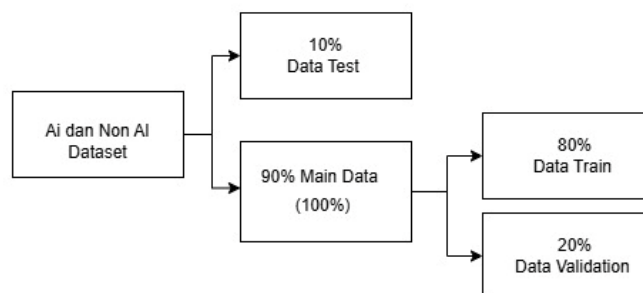
Dataset yang telah dipilih kemudian dibagi ke dalam tiga subset utama, yaitu *training set*, *validation set*, dan *testing set*, untuk mendukung proses pembelajaran terawasi (*supervised learning*).

Pada tahap awal, dataset dibagi menjadi **90% data pelatihan** dan **10% data pengujian (testing set)**. Selanjutnya, data pelatihan dibagi kembali menjadi **80% data training** dan **20% data validation**. Berdasarkan skema tersebut, diperoleh distribusi data sebagai berikut:

- **4.320 citra (72%)** sebagai *training set*
- **1.080 citra (18%)** sebagai *validation set*
- **600 citra (10%)** sebagai *testing set*

Data *training* digunakan untuk melatih model melalui proses optimasi parameter, sedangkan data *validation* digunakan untuk memantau kinerja model selama pelatihan. Data *testing* tidak dilibatkan dalam proses pelatihan dan hanya digunakan untuk evaluasi akhir guna mengukur kemampuan generalisasi model.

Seluruh subset data mempertahankan keseimbangan jumlah kelas REAL dan FAKE untuk menghindari bias klasifikasi. Alur pembagian data ditunjukkan pada **Gambar 4**.



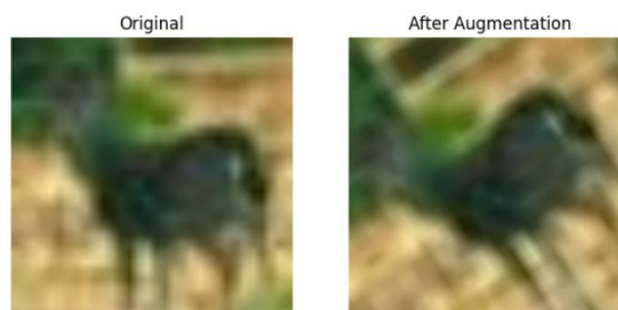
Gambar 4. Diagram Split Data

2.3 Augmentasi Data

Augmentasi data digunakan untuk meningkatkan variasi data pelatihan dan kemampuan generalisasi model, khususnya pada kondisi keterbatasan jumlah data[11]. Dalam penelitian ini, pemilihan teknik augmentasi dilakukan secara hati-hati karena tugas klasifikasi berfokus pada deteksi anomali tekstural pada citra hasil generasi AI.

Teknik augmentasi yang diterapkan dibatasi pada transformasi geometris ringan, yaitu rotasi acak dengan rentang -15° hingga $+15^\circ$, pergeseran horizontal dan vertikal hingga 10%, shear transformation dengan nilai maksimum 0,1, serta horizontal flipping secara acak. Seluruh augmentasi diterapkan hanya pada data pelatihan.

Augmentasi berbasis warna dan tekstur, seperti color jittering, brightness, dan contrast adjustment, tidak digunakan karena berpotensi mengubah karakteristik tekstur dan mengaburkan artefak visual halus yang menjadi indikator utama citra sintetis. Pendekatan augmentasi ini bertujuan menjaga informasi diskriminatif penting sekaligus meningkatkan robustnes model[12]. Seluruh proses augmentasi diterapkan secara *on-the-fly* hanya pada data *training*, sehingga setiap epoch dapat menerima variasi citra yang berbeda. Pendekatan ini memungkinkan model untuk mempelajari fitur yang lebih robust terhadap variasi posisi, orientasi, dan skala objek. Contoh hasil augmentasi ditampilkan pada **Gambar 5**.



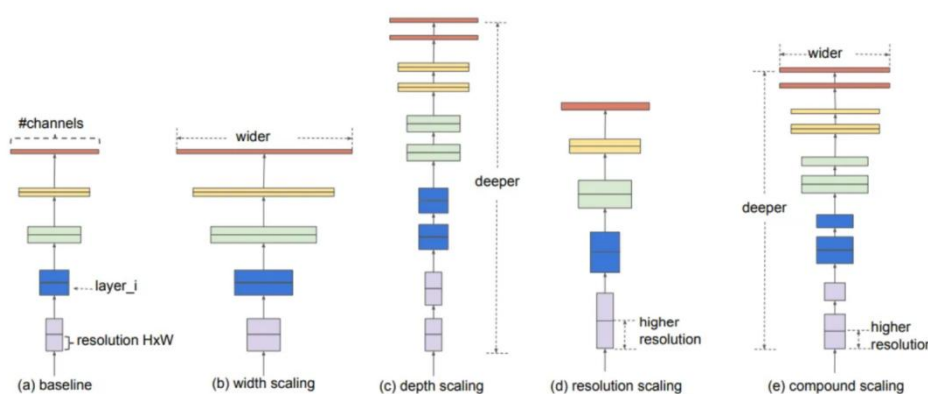
Gambar 5. Hasil Image Augmentation

2.4 Perancangan Model

Model Convolutional Neural Network (CNN) yang digunakan pada penelitian ini adalah EfficientNet-B2, sebuah arsitektur yang dikembangkan menggunakan metode penskalaan majemuk (compound scaling)[13]. Metode ini menyeimbangkan penskalaan pada tiga dimensi jaringan secara

bersamaan: kedalaman (depth), lebar (width), dan resolusi input gambar (resolution)[14]. Pendekatan ini menghasilkan model yang lebih efisien secara komputasi karena dapat mencapai akurasi yang tinggi dengan jumlah parameter yang lebih sedikit jika dibandingkan dengan CNN konvensional seperti ResNet. Efisiensi ini membuat EfficientNet-B2 sangat cocok untuk tugas klasifikasi citra yang memerlukan analisis tekstur mendetail, seperti deteksi anomali pada gambar generatif AI.

Secara arsitektural, EfficientNet-B2 dibangun di atas jaringan dasar (baseline) EfficientNet-B0 yang ditemukan melalui pencarian arsitektur neural (Neural Architecture Search)[13]. Inti dari arsitektur ini adalah blok MBConv (Mobile Inverted Bottleneck Convolution) yang dioptimalkan dengan mekanisme Squeeze-and-Excitation untuk secara dinamis menekankan fitur yang lebih relevan[14]. Struktur lengkap arsitektur EfficientNet-B2 yang digunakan dalam penelitian ini ditunjukkan pada **Gambar 6**.



Gambar 6. EfiicientNetB2 Arstitektur

2.5 Optimisasi dan Pelatihan Model

Pada tahap pelatihan model, digunakan optimizer Adam yang mampu memperbarui bobot secara adaptif sehingga proses pembelajaran menjadi lebih stabil dan efisien [15][16]. Beberapa penelitian juga menunjukkan bahwa Adam memiliki kemampuan konvergensi yang baik pada model deep learning dibandingkan metode optimisasi lainnya [16][17].

Proses pelatihan dilakukan secara bertahap (staged training). Tahap awal menggunakan Stochastic Gradient Descent (SGD) dengan learning rate 0,01 dan momentum 0,9 selama 3 epoch sebagai baseline, dengan seluruh lapisan EfficientNet-B2 dalam kondisi frozen. Selanjutnya, pelatihan dilanjutkan menggunakan Adam dengan learning rate 1×10^{-4} selama 10 epoch.

Pada tahap fine-tuning, sebanyak 20 lapisan terakhir dari model dibuka (trainable) dan dilatih menggunakan Adam dengan learning rate 1×10^{-5} selama 5 epoch untuk melakukan penyesuaian parameter secara lebih halus.

Pelatihan menggunakan fungsi loss binary crossentropy dan metrik accuracy, dengan ukuran batch size 32 serta input citra berukuran 260×260 piksel. Untuk mengurangi overfitting, diterapkan Early Stopping dengan patience 5 epoch berdasarkan validation loss, serta ReduceLROnPlateau dengan faktor penurunan learning rate sebesar 0,5 dalam 3 epoch tanpa peningkatan performa.

Setelah proses pelatihan selesai, model diharapkan telah mampu mempelajari pola representatif dari data latih dan dapat digunakan untuk melakukan klasifikasi pada data baru [18].

2.6 Evaluasi Model

Pada tahap evaluasi, penelitian ini menggunakan confusion matrix untuk menilai kemampuan model dalam membedakan dua kelas, yaitu Gambar AI dan Gambar Non-AI. Confusion matrix yang digunakan merupakan matriks ordo 2×2 yang menampilkan jumlah prediksi benar maupun salah untuk masing-masing kelas, sebagaimana ditunjukkan pada **Tabel 1**. Matriks ini membantu mengidentifikasi sejauh mana model mampu mengenali citra sesuai label aslinya.

Table 1. Contoh Confusion Matrix 2x2

Actual/Predicted	Fake	Real
Fake	TP	FN
Real	FP	TN

Selain itu, penelitian ini juga menggunakan classification report untuk memberikan analisis yang lebih mendalam terkait performa model pada data testing. Laporan ini menyajikan metrik evaluasi seperti precision, recall, dan f1-score[6]. Precision menunjukkan proporsi prediksi positif yang benar, recall menggambarkan kemampuan model dalam menemukan data positif secara tepat, sedangkan f1-score merupakan rata-rata harmonik yang menyeimbangkan keduanya. Evaluasi ini memberikan gambaran komprehensif mengenai efektivitas model dalam mengklasifikasi gambar AI dan Non-AI secara akurat.

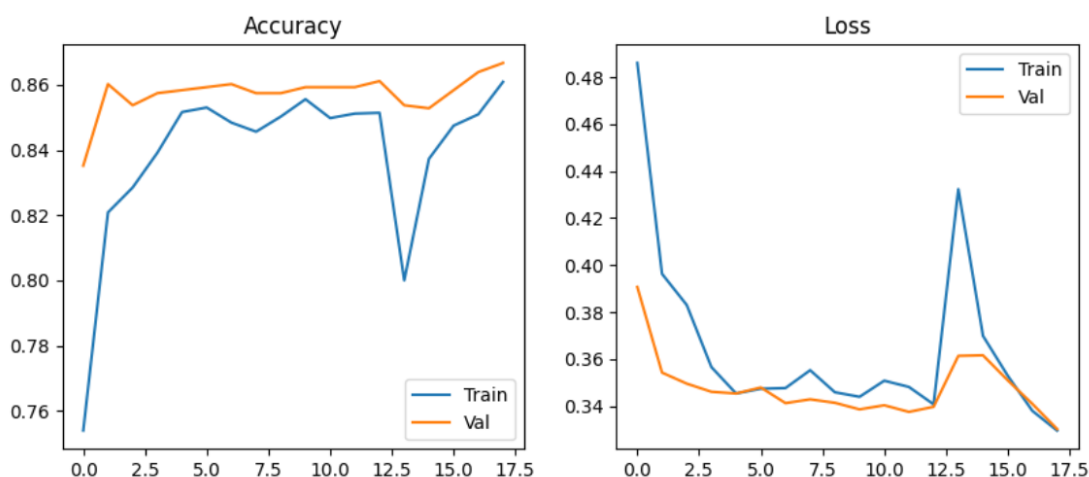
3. HASIL DAN PEMBAHASAN

3.1. Hasil Pelatihan Model

Hasil pelatihan model menunjukkan bahwa arsitektur EfficientNet-B2 mampu mempelajari representasi fitur visual yang efektif untuk membedakan citra FAKE dan REAL. Pada tahap awal (baseline), model menghasilkan akurasi validasi sebesar 85,37%, yang menunjukkan bahwa fitur yang diekstraksi dari model pretrained sudah cukup relevan untuk tugas klasifikasi biner pada dataset yang digunakan.

Setelah dilakukan proses optimisasi menggunakan Adam optimizer, performa model mengalami peningkatan dengan akurasi validasi mencapai 86,11%. Peningkatan ini juga diikuti dengan penurunan validation loss yang lebih stabil, sebagaimana terlihat pada **Gambar 7**, yang menunjukkan bahwa proses pembelajaran berlangsung lebih terkontrol.

Pada tahap fine-tuning, model mencapai performa terbaik dengan akurasi validasi sebesar 86,67% dan nilai validation loss sebesar 0,3302. Peningkatan ini mengindikasikan bahwa penyesuaian bobot pada lapisan yang lebih dalam memungkinkan model menangkap pola yang lebih kompleks.



Gambar 7. Grafik training vs validation accuracy

Berdasarkan grafik accuracy, terlihat bahwa kurva training dan validation memiliki tren yang relatif sejalan tanpa adanya gap yang signifikan. Hal ini menunjukkan bahwa model telah mencapai kondisi konvergen dan mampu belajar secara konsisten sepanjang proses pelatihan.

Sementara itu, pada grafik loss, nilai training loss dan validation loss cenderung menurun dan stabil di akhir epoch, meskipun terdapat sedikit fluktuasi pada beberapa titik. Tidak adanya perbedaan yang mencolok antara kedua kurva menunjukkan bahwa model tidak mengalami overfitting yang signifikan.

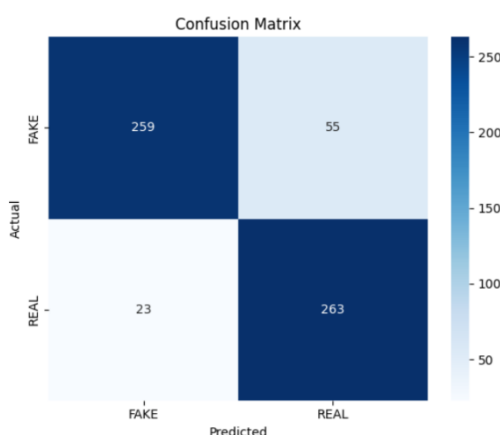
Secara keseluruhan, tren peningkatan performa dari tahap baseline hingga fine-tuning, serta pola kurva yang stabil, menunjukkan bahwa strategi pelatihan bertahap yang diterapkan mampu meningkatkan kualitas model secara konsisten. Hal ini mengindikasikan bahwa kombinasi transfer learning dan fine-tuning merupakan pendekatan yang efektif dalam tugas klasifikasi citra berbasis tekstur.

3.2. Evaluasi Model

Tahap evaluasi dilakukan menggunakan data testing yang terdiri dari 600 citra, yang sepenuhnya terpisah dari data pelatihan dan validasi. Evaluasi ini bertujuan untuk mengukur performa akhir model setelah melalui proses pelatihan dan fine-tuning, serta untuk menilai kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Dua metode evaluasi utama digunakan dalam penelitian ini, yaitu confusion matrix dan classification report, yang merupakan pendekatan standar dalam evaluasi model klasifikasi biner.

Confusion matrix berordo 2×2 digunakan untuk menganalisis distribusi prediksi model pada dua kelas, yaitu FAKE dan REAL. Matriks ini terdiri dari empat komponen utama, yaitu True Positive (TP), False Negative (FN), False Positive (FP), dan True Negative (TN), yang masing-masing merepresentasikan prediksi benar dan salah terhadap label aktual. Analisis confusion matrix memungkinkan identifikasi pola kesalahan model secara lebih rinci, sehingga dapat diketahui kecenderungan kesalahan klasifikasi yang terjadi pada masing-masing kelas.

Hasil pengujian menunjukkan bahwa model mencapai akurasi testing sebesar 87,12% dengan nilai loss sebesar 0,2897, yang mengindikasikan performa generalisasi yang baik. Berdasarkan classification report, pada kelas FAKE model memperoleh nilai precision sebesar 0,92 dan recall sebesar 0,82, sedangkan pada kelas REAL diperoleh precision sebesar 0,83 dan recall sebesar 0,92. Nilai f1-score pada kedua kelas relatif seimbang, yaitu 0,87, yang menunjukkan bahwa model mampu menjaga keseimbangan antara presisi dan sensitivitas dalam mendeteksi kedua kelas.



Gambar 8. Confusion Matrix

Table 2. Confusion Matrix 2X2

Predicted/Actual	Fake	Real
Fake	259	55
Real	23	263

Confusion matrix yang ditunjukkan pada **Gambar 8** serta dirangkum pada **Tabel 2** memperlihatkan bahwa model cenderung lebih sering melakukan kesalahan berupa False Negative, yaitu mengklasifikasikan citra FAKE sebagai REAL, dibandingkan kesalahan sebaliknya. Fenomena ini tercermin dari nilai recall kelas FAKE yang lebih rendah dibandingkan kelas REAL. Meskipun demikian, jumlah prediksi benar pada kedua kelas tetap mendominasi, yang menunjukkan bahwa model memiliki kemampuan yang stabil dalam mengenali pola visual baik pada citra generatif berbasis AI maupun citra asli.

3.3. Pembahasan

Hasil evaluasi menunjukkan bahwa model memiliki kecenderungan lebih sulit dalam mendeteksi citra FAKE dibandingkan citra REAL. Hal ini ditunjukkan oleh jumlah False Negative yang relatif lebih tinggi, yang mengindikasikan bahwa citra sintetis masih sering diklasifikasikan sebagai citra nyata.

Fenomena ini dipengaruhi oleh karakteristik dataset, di mana citra FAKE dihasilkan menggunakan model generatif yang mampu menyerupai distribusi visual citra asli. Perbedaan antara kedua kelas menjadi sangat halus, khususnya pada aspek tekstural. Selain itu, resolusi citra yang relatif rendah (32×32 piksel) membatasi kemampuan model dalam menangkap detail lokal, sehingga model cenderung mengandalkan pola global yang kurang diskriminatif terhadap artefak sintetis.

Strategi augmentasi yang digunakan juga turut berpengaruh. Meskipun augmentasi geometris membantu meningkatkan variasi data, pembatasan pada transformasi tertentu dapat mengurangi keberagaman pola artefak yang dipelajari, sehingga sensitivitas model terhadap citra FAKE berkualitas tinggi menjadi terbatas.

Berdasarkan hasil evaluasi, model yang diusulkan memperoleh akurasi sebesar 87,12%, yang menunjukkan performa cukup baik dalam membedakan citra REAL dan FAKE. Namun, jika dibandingkan dengan penelitian oleh Bird dan Lotfi [9] yang mencapai akurasi 92,98% pada dataset serupa, performa model pada penelitian ini masih berada di bawah hasil tersebut.

Perbedaan ini kemungkinan disebabkan oleh beberapa faktor, seperti keterbatasan jumlah data yang digunakan, eksplorasi arsitektur model yang lebih terbatas, serta strategi hyperparameter tuning yang tidak sekomprehensif penelitian terdahulu. Meskipun demikian, pendekatan yang digunakan dalam penelitian ini tetap menunjukkan performa yang kompetitif dengan konfigurasi model yang relatif lebih sederhana dan efisien.

Dengan demikian, penggunaan EfficientNet-B2 yang dikombinasikan dengan strategi pelatihan bertahap terbukti cukup efektif untuk tugas deteksi citra generatif. Peningkatan performa masih memungkinkan dilakukan, terutama melalui penggunaan dataset dengan resolusi lebih tinggi, variasi data yang lebih luas, serta optimasi model yang lebih mendalam.

4. KESIMPULAN

Penelitian ini berhasil membangun model klasifikasi citra berbasis EfficientNet-B2 untuk membedakan citra AI-generated dan non-AI secara efektif. Pemanfaatan arsitektur pretrained sebagai feature extractor, yang dikombinasikan dengan strategi optimisasi bertahap menggunakan Adam optimizer, terbukti mampu meningkatkan performa model secara konsisten. Hasil evaluasi menunjukkan bahwa model mencapai akurasi sebesar 87,12% pada data pengujian, dengan nilai precision, recall, dan f1-score yang relatif seimbang pada kedua kelas. Capaian tersebut mengindikasikan bahwa model memiliki kemampuan klasifikasi yang reliabel serta tingkat generalisasi yang baik terhadap data baru.

Secara keseluruhan, hasil penelitian ini membuktikan bahwa kombinasi EfficientNet-B2 dan pendekatan fine-tuning yang terkontrol merupakan solusi yang efektif untuk mendeteksi citra hasil generatif AI. Model yang diusulkan mampu mengekstraksi fitur visual yang diskriminatif dan menunjukkan stabilitas performa pada seluruh tahap pelatihan dan evaluasi.

Meskipun penelitian ini menunjukkan performa yang cukup baik, beberapa keterbatasan masih perlu diperhatikan. Dataset yang digunakan memiliki resolusi citra relatif rendah (32×32 piksel), sehingga artefak visual khas citra generatif berpotensi tidak terekam secara optimal, khususnya pada citra AI dengan kualitas tinggi. Selain itu, sumber citra FAKE masih terbatas pada generator tertentu dan teknik augmentasi yang digunakan bersifat konservatif, sehingga variasi karakteristik citra sintetis yang dipelajari model belum sepenuhnya merepresentasikan kondisi dunia nyata. Oleh karena itu, penggunaan dataset beresolusi lebih tinggi, sumber generator yang lebih beragam, serta strategi pelatihan yang lebih komprehensif berpotensi meningkatkan performa model di masa mendatang.

Sebagai pengembangan lebih lanjut, terdapat beberapa saran yang dapat dipertimbangkan pada penelitian berikutnya. Pertama, perluasan dataset dengan jumlah data yang lebih besar dan variasi yang lebih beragam, termasuk citra hasil dari berbagai image generator seperti Midjourney, Stable Diffusion, dan Leonardo AI, diharapkan dapat meningkatkan kemampuan generalisasi model terhadap beragam karakteristik citra AI. Kedua, eksperimen dengan arsitektur model lain seperti EfficientNetV2, ConvNeXt, atau Vision Transformer (ViT) dapat dilakukan sebagai studi perbandingan untuk

mengevaluasi potensi peningkatan performa dan efisiensi komputasi. Ketiga, penambahan teknik augmentasi data yang lebih variatif, seperti CutMix, MixUp, dan Random Erasing, dapat diterapkan untuk mengurangi risiko overfitting serta meningkatkan ketahanan model terhadap variasi citra di dunia nyata.

Dengan mempertimbangkan saran-saran tersebut, penelitian lanjutan diharapkan mampu menghasilkan sistem deteksi citra AI yang lebih robust, adaptif, dan siap diterapkan pada skala yang lebih luas.

5. DAFTAR PUSTAKA

- [1] A. Borji, "Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2," pp. 1–9, 2022, [Online]. Available: <http://arxiv.org/abs/2210.00586>
- [2] G. P. Kurniawan, S. Z. Shalikhah, H. Shofiat, N. N. Azizah, and Mahmud Mochtar, "Jurnal Tana Mana," *J. Tana Mana*, vol. 2, no. 1, pp. 46–48, 2021.
- [3] T. Z. Alam and J. Haikal, "Dampak Produksi Desain Grafis Pada Penggunaan Teknologi Artificial Intelligence (AI) Dengan Menggunakan Grounded Theory," *J. Seni Nas. Cikini*, vol. 10, no. 1, pp. 15–26, 2024, doi: 10.52969/jsnc.v10i1.265.
- [4] M. S. A. Aria, C. Slamet, and M. D. Firdaus, "Klasifikasi Fake dan Real Menggunakan Vision Transformer dan EfficientNet-B0 pada Gambar Asli dan Generatif AI," *Smatika J.*, vol. 15, no. 01, pp. 179–192, 2025, doi: 10.32664/smatika.v15i01.1531.
- [5] M. S. Maulana, "DETEKSI OSTEOPOROSIS PADA CITRA X-RAY TULANG TANGAN BERDASARKAN FITUR TEKSTUR, GEOMETRI, DAN RUANG WARNA MENGGUNAKAN METODE NAIVE BAYES SKRIPSI Oleh: Muhammad Setya Maulana NIM. 19640026 PROGRAM STUDI FISIKA FAKULTAS SAINS DAN TEKNOLOGI UNIVERSITAS ISLAM," 2023.
- [6] H. K. Maulana, "Penerapan Arsitektur Cnn-Efficientnetb2 Dengan Transfer Learning Pada Klasifikasi Gambar Tokoh Wayang Kulit," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 1, 2025, doi: 10.23960/jitet.v13i1.5626.
- [7] J. Witjaksono, M. Y. Pusadan, Y. Anshori, R. Ardiansyah, and R. Azhar, "Klasifikasi Jenis Batik Bomba Menggunakan Convolutional Neural Network Dengan Arsitektur Efficient-Net B2 (Batik Bomba Sulawesi Tengah)," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.)*, vol. 10, no. 3, pp. 2134–2147, 2025, doi: 10.29100/jipi.v10i3.6191.
- [8] W. O. Mardiana, N. A., & Windari, "G-Tech : Jurnal Teknologi Terapan," *G-Tech J. Teknol. Terap.*, vol. 8, no. 1, pp. 186–195, 2024.
- [9] J. J. Bird, A. Lotfi, and S. Member, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," *IEEE Access*, vol. 12, no. February, pp. 15642–15650, 2024, doi: 10.1109/ACCESS.2024.3356122.
- [10] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," 2009.
- [11] S. Kumar and P. Asiamah, "Enhancing Image Classification with Augmentation: Data Augmentation Techniques for Improved Image Classification".
- [12] X. F. Liu, G. Karagoz, and N. Meratnia, "Analyzing the Impact of Data Augmentation on the Explainability of Deep Learning-Based Medical Image Classification," pp. 1–28, 2025.
- [13] Viso.ai, "EfficientNet: Boost CNN Accuracy with Less Computer." [Online]. Available: <https://viso.ai/deep-learning/efficientnet/>
- [14] GeeksforGeeks, "Efficientnet Architecture." [Online]. Available: <https://www.geeksforgeeks.org/computer-vision/efficientnet-architecture/>
- [15] S. Pariyasto, V. A. Warongan, S. Tinggi, I. Kesehatan, and M. Sejati, "Optimalisasi klasifikasi citra medis menggunakan cnn dan adam optimizer dengan paramater minimum," vol. 4307, no. 1, pp. 691–701, 2025.

- [16] S. Bashetty, K. Raja, S. Adepu, and A. Jain, “Optimizers in Deep Learning : A Comparative,” no. December, 2022.
- [17] I. Journal and O. F. Science, “A Comparative Study of Optimization,” pp. 803–810, 2025.
- [18] M. Reyad and A. M. Sarhan, “A modified Adam algorithm for deep neural network optimization,” *Neural Comput. Appl.*, vol. 35, no. 23, pp. 17095–17112, 2023, doi: 10.1007/s00521-023-08568-z.