SMATIKA : STIKI Informatika Jurnal
Vol. 15, No. 1, Juni 2025, pp. 179~192
ISSN: 2087-0256, e-ISSN: 2580-6939

# Klasifikasi Fake dan Real Menggunakan *Vision Transformer* dan EfficientNet-B0 pada Gambar Asli dan Generatif AI

**M. Syahrul Anwar Aria[1*]**
**Cepy Slamet[2]**
**Muhammad Deden Firdaus[3]**

[1,2,3]Teknik Informatika, UIN Sunan Gunung Djati Bandung, Jl. A.H. Nasution No. 105 Cibiru Kota Bandung 40614, Indonesia
[1]syhrlanwr@gmail.com, [2]cepy_lucky@uinsgd.ac.id, [3]deden@uinsgd.ac.id

***Penulis Korespondensi:**
M. Syahrul Anwar Aria
syhrlanwr@gmail.com

**Abstrak**

Kemajuan teknologi kecerdasan buatan (AI) telah memungkinkan penciptaan gambar sintetis yang menyerupai gambar asli, menimbulkan tantangan dalam mendeteksi dan mengklasifikasikan gambar tersebut. Penelitian ini bertujuan untuk mengembangkan model klasifikasi berbasis EfficientNet-B0 dan Vision Transformer (ViT) untuk membedakan gambar asli dan gambar yang dihasilkan oleh AI generatif. Data yang digunakan terdiri dari 30.401 gambar asli dari dataset MSCOCO 2017 dan 30.401 gambar hasil AI generatif dari SyntheticEye AI-Generated Images Dataset di Kaggle. Hasil penelitian menunjukkan bahwa model ViT mencapai akurasi 98% dan EfficientNet-B0 mencapai akurasi 96% dalam mengklasifikasikan gambar. Kesimpulan dari penelitian ini adalah bahwa kedua model memiliki potensi besar dalam mendeteksi manipulasi media digital, dengan ViT menunjukkan performa yang lebih unggul. Implikasi praktis dari penelitian ini adalah pengembangan teknologi yang lebih canggih untuk mendeteksi gambar generatif, yang dapat digunakan dalam berbagai aplikasi nyata seperti keamanan digital dan verifikasi media.
**Kata Kunci: CNN; EfficientNet; Gambar AI Generatif; Klasifikasi gambar; Pemrosesan gambar; Vision Transformer**

***Abstract***

*Advances in artificial intelligence (AI) technology have enabled the creation of synthetic images that resemble real images, posing challenges in detecting and classifying such images. This study aims to develop an EfficientNet-B0 and Vision Transformer (ViT) based classification model to distinguish between real images and images generated by generative AI. The data used consists of 30,401 original images from the MSCOCO 2017 dataset and 30,401 generative AI-generated images from the SyntheticEye AI-Generated Images Dataset on Kaggle. The results showed that the ViT model achieved 98% accuracy and EfficientNet-B0 achieved 96% accuracy in classifying the images. The conclusion of this research is that both models have great potential in detecting digital media manipulation, with ViT showing superior performance. The practical implication of this research is the development of more advanced technologies for detecting generative images, which can be used in various real applications such as digital security and media verification.*
***Keywords: AI-Generated Image; CNN; EfficientNet; Image classification; Image processing; Vision Transformer.***

## 1. Pendahuluan

Kemajuan teknologi kecerdasan buatan *(Artificial Intelligence/AI)* telah menghasilkan dampak signifikan dalam berbagai bidang kehidupan. Salah satu inovasi yang menarik sekaligus kontroversial adalah kemampuan AI untuk menghasilkan gambar sintetis yang menyerupai gambar asli. Andrew Ng, seorang ahli AI terkemuka, menyatakan, *"AI is the new electricity"*, mencerminkan potensi besar AI sebagai penggerak utama inovasi teknologi [1]. Namun, perkembangan ini juga membawa tantangan besar, terutama dalam membedakan antara gambar asli dan gambar yang dihasilkan oleh AI generatif. Hal ini menimbulkan kekhawatiran, terutama terkait penyalahgunaan teknologi ini untuk kejahatan digital, penyebaran informasi palsu, dan pencemaran nama baik.

Permasalahan ini menjadi semakin nyata dengan insiden yang melibatkan foto-foto viral yang ternyata merupakan hasil generasi AI, seperti kasus foto mesum yang diklaim sebagai Taylor Swift pada awal 2024. Setelah investigasi, gambar tersebut diketahui dihasilkan oleh AI [2]. Selain itu, dalam kompetisi fotografi seperti 1839 Awards, para juri kesulitan membedakan antara karya manusia dan gambar hasil AI, menunjukkan betapa sulitnya tantangan ini [3].

Studi sebelumnya menunjukkan bahwa akurasi manusia dalam membedakan gambar asli dan gambar generatif hanya mencapai 61% dalam percobaan tertentu, jauh di bawah proyeksi yang diharapkan sebesar 85% [4]. Data ini menunjukkan adanya kebutuhan mendesak untuk mengembangkan teknologi yang lebih canggih dalam mendeteksi dan mengklasifikasikan gambar generatif.

Beberapa penelitian terdahulu telah mencoba mengatasi tantangan ini. Suatu penelitian telah model CNN dengan 6 lapisan *convolutional* dan 3 *max pooling*, mencapai akurasi 91% untuk mendeteksi wajah asli dan deepfake [5]. Penelitian lain memanfaatkan arsitektur Transformer BEiT untuk klasifikasi gambar generatif dan gambar buatan tangan, namun hanya mencapai akurasi 80% [6]. Ada penelitian yang mengembangkan CNN untuk mendeteksi gambar sintetis pada dataset CIFAKE dengan akurasi 92.98% [7]. Pendekatan terbaru dengan arsitektur Hybrid VGG16-CNN mencatatkan akurasi 94% untuk mendeteksi konten deepfake [8].

Arsitektur EfficientNet, yang diperkenalkan oleh Tan dan Le, menawarkan solusi yang lebih efisien melalui teknik *compound scaling*. Pendekatan ini secara bersamaan mengoptimalkan ukuran jaringan, kedalaman, dan resolusi gambar, menjadikannya pilihan ideal untuk tugas klasifikasi citra yang kompleks [9]. Penelitian sebelumnya telah menunjukkan keberhasilan EfficientNet dalam berbagai aplikasi, seperti deteksi penyakit pada daun padi dengan akurasi hingga 98.91% [10], klasifikasi motif batik Papua dengan akurasi 90% [11], dan deteksi Bahasa Isyarat Indonesia (BISINDO) dengan akurasi 99.24% pada perangkat Android [12]. Selain itu, penelitian oleh Adam dan Santoso menunjukkan penerapan arsitektur EfficientNet untuk klasifikasi penerima bantuan langsung tunai di DKI Jakarta dengan akurasi validasi 95.03% [13]. Penelitian lain oleh Fajrina et al.menggunakan EfficientNet-B0 untuk klasifikasi leukemia tipe Acute Lymphoblastic Leukemia (ALL) dengan akurasi hingga 98.48% [14]. Perdani et al. juga berhasil menerapkan EfficientNet dalam klasifikasi glaukoma berdasarkan citra fundus mata dengan hasil akurasi, presisi, recall, dan F1-Score mencapai 1,0000 [15]. Temuan ini memperkuat relevansi dan efektivitas arsitektur EfficientNet untuk berbagai tugas klasifikasi citra.

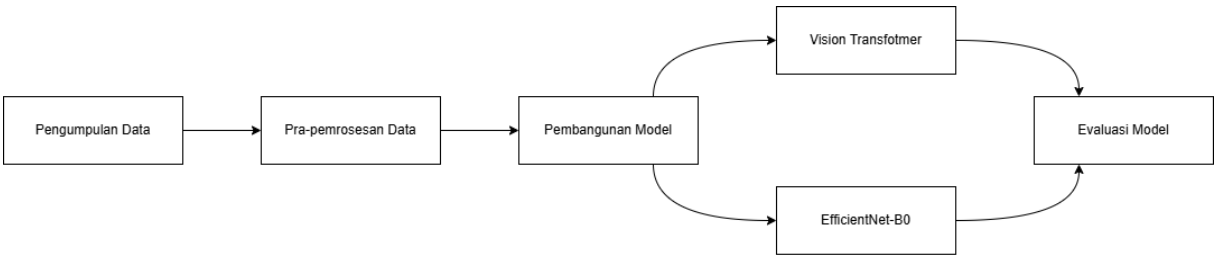
Selain EfficientNet, arsitektur Transformer, yang awalnya dikembangkan untuk pemrosesan bahasa alami, telah diadaptasi untuk tugas klasifikasi gambar. Vision Transformer (ViT), menggunakan pendekatan berbasis patch dan mekanisme *self-attention* untuk memahami hubungan spasial dalam citra [16]. Studi terkini menunjukkan potensi besar ViT dalam klasifikasi citra. Sebagai contoh, menggunakan ViT untuk klasifikasi 15 jenis ikan koi, mencapai akurasi 89%, yang lebih tinggi dibandingkan pendekatan CNN sebelumnya [17]. Penelitian lain menunjukkan bahwa ViT-L/16-in21k mampu mengklasifikasikan tingkat kematangan pisang dengan akurasi 91.61%, mengungguli model CNN [18]. Penelitian lain menggunakan arsitektur ViT-B16 untuk mengklasifikasikan penyakit daun padi, seperti Bercak Coklat, Blast, Hawar Daun Bakteri, dan Tungro, mencapai akurasi 96% [19]. Penelitian lain mengaplikasikan ViT dalam deteksi Covid-19 dari citra x-ray, membandingkannya dengan berbagai arsitektur seperti ResNet50, dan menemukan bahwa pendekatan berbasis Transformer memberikan hasil kompetitif meskipun mengalami tantangan overfitting [20]. Selain itu, Sukandar et al. menunjukkan efektivitas hibrida CNN-ViT dalam klasifikasi tumor otak, mencapai akurasi 94% dengan optimasi menggunakan Adam [21].

Kompleksitas dataset gambar generatif yang melibatkan variasi gaya visual dan tingkat kemiripan tinggi dengan gambar asli memerlukan solusi yang lebih inovatif. Oleh karena itu, penelitian ini

bertujuan mengembangkan model klasifikasi berbasis EfficientNet-B0 dan membandingkannya dengan ViT untuk klasifikasi gambar asli dan generatif. Pendekatan ini akan mengevaluasi efisiensi dan akurasi masing-masing model. Dengan inovasi ini, penelitian diharapkan dapat memberikan kontribusi dalam mengatasi tantangan klasifikasi gambar generatif.

## 2. Metode Penelitian

Penelitian ini dilakukan dengan tujuan untuk mengembangkan model klasifikasi berbasis EfficientNet-B0 dan Vision Transformer (ViT) untuk membedakan gambar asli dan gambar yang dihasilkan oleh kecerdasan buatan generatif. Berikut adalah tahapan penelitian ini seperti yang ditunjukkan pada Gambar 1:




***Gambar 1.*** *Tahapan Penelitian*

Data yang digunakan dalam penelitian ini terdiri dari dua jenis gambar, yaitu gambar asli dan gambar hasil kecerdasan buatan generatif. Gambar asli diambil dari dataset MSCOCO 2017 yang berisi berbagai jenis gambar yang menggambarkan objek dan adegan dari dunia nyata [22]. Untuk gambar hasil kecerdasan buatan generatif, digunakan SyntheticEye AI-Generated Images Dataset yang tersedia di platform Kaggle [23]. Dataset ini berisi 30.401 gambar yang dihasilkan menggunakan AI generatif seperti Stable Diffusion, OpenJourney, dan Min-Dalle [23]. Total data yang digunakan dalam penelitian ini adalah 60.802 gambar, dengan rincian 30.401 gambar asli dan 30.401 gambar hasil AI generatif. Gambar 2 menunjukkan beberapa sampel dari dataset yang digunakan dalam penelitian ini.


***Gambar 2.*** *Sampel data*

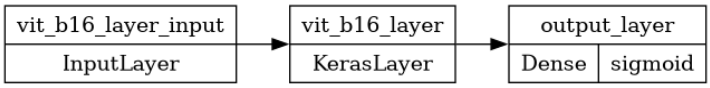
Pra pemrosesan data dilakukan untuk menyiapkan dataset sebelum digunakan dalam proses pelatihan model. Tahapan pra-pemrosesan ini meliputi, semua gambar diubah ukurannya menjadi 224x224 piksel untuk menyesuaikan dengan input yang diperlukan oleh arsitektur EfficientNet-B0. Dataset dibagi menjadi tiga bagian, yaitu data pelatihan, data validasi dan data pengujian dengan rasio 60:20:20 seperti yang ada pada Tabel 1. Sebanyak 36.480 gambar

digunakan untuk pelatihan, 12.160 gambar digunakan untuk validasi, dan 12162 gambar digunakan untuk pengujian

***Tabel 1.*** *Pembagian data pelatihan, data validasi, dan data pengujian*

| | *Train* | *Validation* | *Test* | *Total* |
|---|---|---|---|---|
| *Fake* | 18240 | 6080 | 6081 | 30401 |
| *Real* | 18240 | 6080 | 6081 | 30401 |
| ***Total*** | **36480** | **12160** | **12162** | **60802** |

Terdapat dua model yang dikembangkan dalam penelitian ini, yaitu model Vision Transformer (ViT) dan model berbasis EfficientNet-B0. Kedua model ini dilatih untuk mengklasifikasikan gambar asli dan gambar hasil AI generatif. Pada arsitektur pertama, Vision Transformer (ViT) diterapkan menggunakan pre-trained model ViT-B/16 yang diimpor melalui TensorFlow Hub. Gambar 3 menggambarkan arsitektur Vision Transformer yang digunakan dalam penelitian ini. Model ini menggunakan mekanisme self-attention untuk memahami hubungan global dalam citra dan telah dilatih sebelumnya pada dataset ImageNet-21k. Bobot pada Vision Transformer layer ditetapkan sebagai tidak dapat dilatih (*trainable=False*) untuk memanfaatkan kemampuan generalisasi bawaan model. Lapisan terakhir berupa lapisan Dense dengan unit tunggal dan fungsi aktivasi sigmoid yang dirancang untuk klasifikasi biner. Model diimplementasikan menggunakan TensorFlow dengan struktur berikut:




***Gambar 3.*** *Arsitektur Vision Transformer*

Sementara itu, pada arsitektur kedua, *EfficientNet-B0* digunakan sebagai *base model* dengan bobot awal dari *ImageNet*. Model ini tidak menyertakan lapisan *fully connected* pada bagian atas (*include_top=False*) untuk memberikan fleksibilitas pada desain lapisan akhir. Selanjutnya, dilakukan *fine-tuning* pada 20 lapisan terakhir model untuk meningkatkan kemampuan generalisasi terhadap dataset spesifik. Bagian akhir model terdiri atas lapisan *Global Average Pooling* untuk mereduksi dimensi data, diikuti oleh lapisan *Dense* dengan 64 unit dan fungsi aktivasi *ReLU*. Untuk mencegah overfitting, lapisan *Dropout* dengan tingkat 50% diterapkan sebelum lapisan output. Lapisan output menggunakan unit tunggal dengan fungsi aktivasi *sigmoid* untuk menghasilkan nilai probabilitas biner. Gambar 4 menunjukkan struktur arsitektur EfficientNet yang diterapkan dalam penelitian ini. Model ini diimplementasikan dengan pustaka *TensorFlow* dan dirancang untuk menerima masukan gambar berukuran 224x224 piksel. Berikut adalah struktur implementasinya:


***Gambar 4.*** *Arsitektur EfficientNet-B0*

Kedua model dilatih menggunakan *optimizer* Adam dengan *learning rate* 0.001. Fungsi *loss* yang digunakan adalah *binary_crossentropy* karena tugas klasifikasi bersifat biner. Pelatihan dilakukan selama 10 *epoch* dengan ukuran *batch* 16. Berikut adalah paramater training yang digunakan seperti yang ditunjukkan oleh Tabel 2.

***Tabel 2.*** *Parameter Training*

| Nama | Parameter |
|---|---|
| *Optimizer* | Adam |
| *Learning rate* | 0.001 |
| *Loss* | *binary_crossentropy* |
| *Batch size* | 16 |
| *Epoch* | 10 |

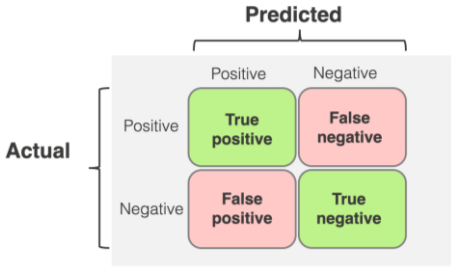
Tabel 3 merupakan spesifikasi dari Google Colab yang akan digunakan untuk melatih model. Model akan dilatih menggunakan Google Colab dengan memanfaatkan GPU Tesla T4 untuk mempercepat proses komputasi. Google Colab menyediakan lingkungan berbasis cloud dengan spesifikasi perangkat keras yang mendukung pemrosesan model deep learning secara efisien.

***Tabel 3.*** *Spesifikasi sistem Google Colab*

| Komponen | Spesifikasi |
|---|---|
| *CPU* | Intel Xeon 2.0 GHz (2 core) |
| *GPU* | NVIDIA Tesla T4 |
| *OS* | *Ubuntu 22.04* |
| *RAM* | 12 GB |
| *Storage* | 100 GB |
| *Bahasa Pemrograman* | Python 3.11 |

Evaluasi model dilakukan dengan mengukur kinerja model berdasarkan beberapa metrik, antara lain akurasi, *precision*, *recall*, *F1-score*, dan *confusion matrix* [24]. *Confusion matrix* digunakan untuk menggambarkan performa model dalam klasifikasi dua kelas, yaitu gambar asli dan gambar hasil AI generatif.

Gambar 5 menampilkan confusion matrix yang digunakan untuk mengevaluasi hasil prediksi model. Matriks ini menunjukkan jumlah prediksi yang benar dan salah untuk masing-masing kelas, termasuk True Positives (TP) dan True Negatives (TN) yang menandakan prediksi yang sesuai dengan label sebenarnya, serta False Positives (FP) dan False Negatives (FN) yang menunjukkan kesalahan prediksi.




***Gambar 5.*** *Confusion Matrix [25]*

True Positive (TP) adalah sampel yang termasuk dalam kelas positif dan diklasifikasikan dengan benar sebagai kelas positif. True Negative (TN) adalah sampel yang termasuk dalam kelas negatif dan diklasifikasikan dengan benar sebagai kelas negatif. False Positive (FP) adalah sampel yang termasuk dalam kelas negatif tetapi diklasifikasikan salah sebagai kelas positif. False Negative (FN) adalah sampel yang termasuk dalam kelas positif tetapi diklasifikasikan salah sebagai kelas negatif.

Rumus yang digunakan untuk menghitung metrik tersebut adalah sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

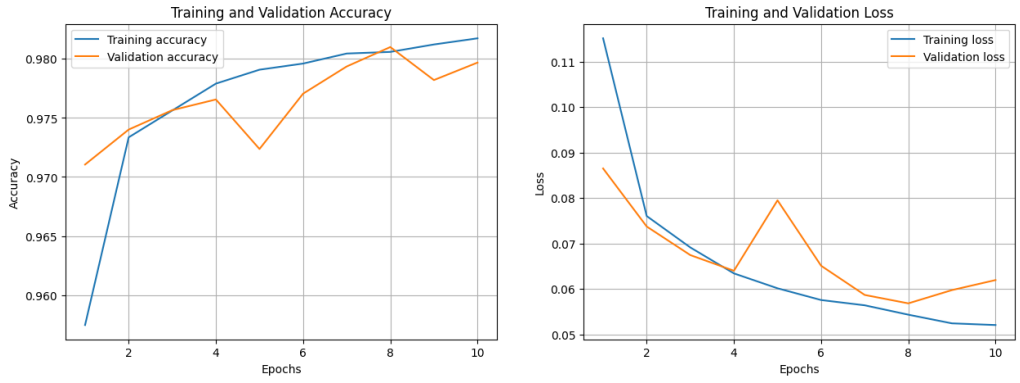
$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

## 3. Hasil

Hasil pelatihan selama 10 epoch dapat dilihat pada Tabel 4. Model menunjukkan peningkatan yang signifikan dalam metrik evaluasi seperti akurasi, precision, dan recall pada setiap epoch. Pada epoch pertama, model mencapai akurasi sebesar 0.9575 dengan loss sebesar 0.1152, yang diikuti dengan peningkatan akurasi hingga mencapai 0.9817 pada epoch terakhir, dengan nilai loss yang menurun menjadi 0.0520. Precision dan recall model juga menunjukkan hasil yang sangat baik, dengan precision mencapai 0.9803 dan recall sebesar 0.9832 pada epoch ke-10. Untuk data validasi, akurasi model meningkat dari 0.9711 pada epoch pertama menjadi 0.9797 pada epoch terakhir, dengan nilai loss yang relatif stabil di sekitar 0.06. Gambar 6 menunjukkan grafik akurasi dan loss untuk model Vision Transformer.

***Tabel 4.*** *Hasil Training model Vision Transformer*

| | *Train* | | | | *Validation* | | | |
|---|---|---|---|---|---|---|---|---|
| *Epoch* | *Accuracy* | *Loss* | *Precision* | *Recall* | *Accuracy* | *Loss* | *Precision* | *Recall* |
| 1 | 0.9575 | 0.1152 | 0.9532 | 0.9621 | 0.9711 | 0.0865 | 0.9677 | 0.9747 |
| 2 | 0.9734 | 0.0760 | 0.9702 | 0.9768 | 0.9740 | 0.0737 | 0.9723 | 0.9758 |
| 3 | 0.9756 | 0.0691 | 0.9735 | 0.9779 | 0.9757 | 0.0674 | 0.9712 | 0.9804 |
| 4 | 0.9779 | 0.0634 | 0.9760 | 0.9799 | 0.9766 | 0.0640 | 0.9698 | 0.9837 |
| 5 | 0.9791 | 0.0601 | 0.9774 | 0.9809 | 0.9724 | 0.0795 | 0.9837 | 0.9607 |
| 6 | 0.9796 | 0.0575 | 0.9737 | 0.9811 | 0.9771 | 0.0651 | 0.9760 | 0.9781 |
| 7 | 0.9805 | 0.0564 | 0.9793 | 0.9816 | 0.9794 | 0.0587 | 0.9788 | 0.9799 |
| 8 | 0.9806 | 0.0543 | 0.9796 | 0.9816 | 0.9810 | 0.0568 | 0.9820 | 0.9799 |
| 9 | 0.9812 | 0.0524 | 0.9798 | 0.9827 | 0.9782 | 0.0597 | 0.9695 | 0.9875 |
| 10 | 0.9817 | 0.0520 | 0.9803 | 0.9832 | 0.9797 | 0.0619 | 0.9705 | 0.9895 |

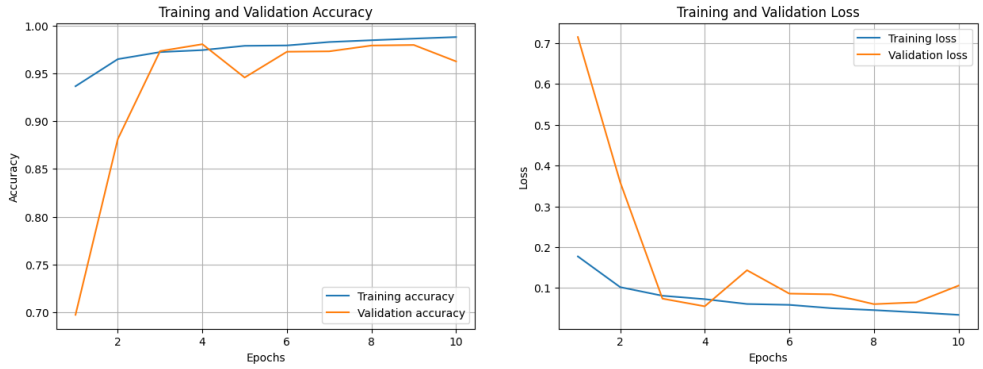



***Gambar 6.*** *Grafik akurasi dan loss model Vision Transformer*

***Tabel 5.*** *Hasil Training model EfficientNetB0*

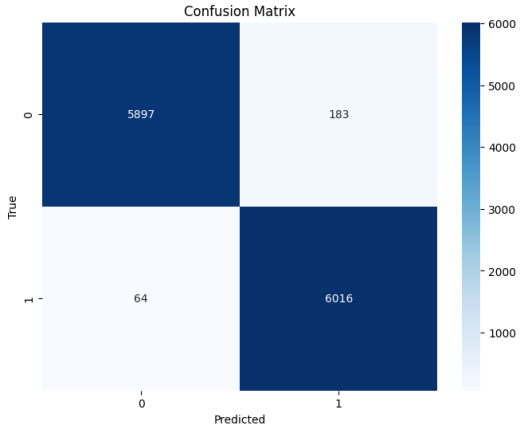
| | *Train* | | | | *Validation* | | | |
|---|---|---|---|---|---|---|---|---|
| *Epoch* | *Accuracy* | *Loss* | *Precision* | *Recall* | *Accuracy* | *Loss* | *Precision* | *Recall* |
| 1 | 0.9154 | 0.2300 | 0.9073 | 0.9278 | 0.6973 | 0.7151 | 0.9781 | 0.4036 |
| 2 | 0.9636 | 0.1058 | 0.9573 | 0.9701 | 0.8812 | 0.3596 | 0.8091 | 0.9977 |
| 3 | 0.9714 | 0.0826 | 0.9674 | 0.9760 | 0.9733 | 0.0738 | 0.9724 | 0.9742 |
| 4 | 0.9757 | 0.0708 | 0.9729 | 0.9785 | 0.9805 | 0.0550 | 0.9785 | 0.9826 |
| 5 | 0.9793 | 0.0569 | 0.9749 | 0.9838 | 0.9456 | 0.1435 | 0.9615 | 0.9285 |
| 6 | 0.9783 | 0.0618 | 0.9752 | 0.9814 | 0.9726 | 0.0861 | 0.9543 | 0.9928 |
| 7 | 0.9834 | 0.0464 | 0.9826 | 0.9840 | 0.9730 | 0.0844 | 0.9607 | 0.9863 |
| 8 | 0.9861 | 0.0414 | 0.9832 | 0.9893 | 0.9791 | 0.0604 | 0.9844 | 0.9737 |
| 9 | 0.9881 | 0.0367 | 0.9864 | 0.9897 | 0.9797 | 0.0647 | 0.9699 | 0.9901 |
| 10 | 0.9889 | 0.0304 | 0.9875 | 0.9903 | 0.9625 | 0.1057 | 0.9368 | 0.9919 |

Pada model EfficientNetB0, hasil pelatihan juga menunjukkan peningkatan performa yang signifikan seperti terlihat pada Tabel 5. Dimulai dengan akurasi 0.9154 pada epoch pertama, model menunjukkan peningkatan yang konsisten hingga mencapai akurasi 0.9889 pada epoch ke-10. Loss pada data pelatihan berkurang drastis dari 0.2300 pada epoch pertama menjadi 0.0304 pada epoch terakhir. Precision dan recall untuk model EfficientNetB0 juga meningkat, dengan precision pada epoch ke-10 mencapai 0.9875 dan recall sebesar 0.9903. Untuk data validasi, meskipun ada fluktuasi kecil dalam nilai akurasi, model menunjukkan akurasi yang relatif tinggi pada 0.9625 pada epoch terakhir, dengan nilai loss 0.1057. Gambar 7 menggambarkan grafik akurasi dan loss untuk model EfficientNetB0.




***Gambar 7.*** *Grafik akurasi dan loss model EfficientnetB0*

## 4. Pembahasan

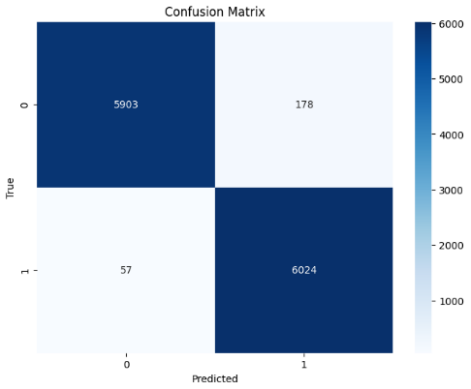



***Gambar 8.*** *Confusion matrix model ViT pada data validasi*

***Tabel 6.*** *Metrik evaluasi model ViT pada data validasi*

| Kelas | *Precision* | *Recall* | *F1-score* |
|---|---|---|---|
| *Fake* | 0.99 | 0.97 | 0.98 |
| *Real* | 0.97 | 0.99 | 0.98 |
| Akurasi | 0.98 | | |

Pada model Vision Transformer, Gambar 8 menunjukkan confusion matrix pada data validasi untuk model ViT, dengan hasil yang sangat baik, yaitu 5897 *true positive (TP)* untuk kategori *fake* dan 6016 TP untuk kategori *real.* Namun, model juga mengalami kesalahan klasifikasi, dengan *64 false positive (FP)* untuk *fake* dan 183 *false negative (FN)* untuk *real.* Hal ini menunjukkan bahwa meskipun model memiliki kemampuan tinggi dalam mengenali kelas yang benar, terdapat beberapa kasus di mana *fake* gambar terkadang diklasifikasikan sebagai *real* dan sebaliknya. Metrik evaluasi yang tercantum pada Tabel 6 menunjukkan *precision* dan *recall* yang sangat baik, dengan *precision* 0.99 untuk kategori *fake* dan *recall* 0.97, serta *precision* 0.97 dan *recall* 0.99 untuk kategori *real*. *F1-score* untuk kedua kelas tersebut juga menunjukkan nilai yang sangat baik, sekitar 0.98, menandakan bahwa model memiliki keseimbangan yang baik antara *precision* dan *recall*.

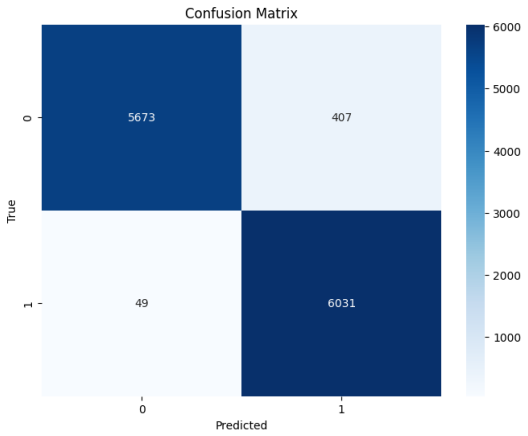



***Gambar 9.*** *Confusion matrix model ViT pada data pengujian*

***Tabel 7.*** *Metrik evaluasi model ViT pada data pengujian*

| Kelas | *Precision* | *Recall* | *F1-score* |
|---|---|---|---|
| *Fake* | 0.99 | 0.97 | 0.98 |
| *Real* | 0.97 | 0.99 | 0.98 |
| Akurasi | 0.98 | | |

Pada Gambar 9, *confusion matrix* untuk dataset pengujian juga menunjukkan hasil yang sangat mirip dengan data validasi. Model mengklasifikasikan 5903 gambar *fake* dengan benar dan 6024 gambar *real*. Namun, ada 178 *false positive* untuk *fake* dan 57 *false negative* untuk *real*. Meskipun kesalahan ini terbilang kecil, mereka tetap perlu dianalisis. Kesalahan false positive dapat terjadi karena beberapa gambar asli memiliki karakteristik visual yang menyerupai pola gambar generatif, misalnya tekstur yang terlalu halus atau detail wajah yang tampak terlalu sempurna akibat proses editing. Sebaliknya, false negative dapat terjadi karena gambar yang dihasilkan oleh AI generatif semakin realistis, dengan pencahayaan, tekstur, dan detail yang menyerupai gambar asli, sehingga model kesulitan membedakannya. Metrik evaluasi pada Tabel 7 menunjukkan bahwa *precision* dan *recall* untuk kedua kelas tetap tinggi, dengan *precision* 0.99 untuk *fake* dan *recall* 0.97, serta *precision* 0.97 dan *recall* 0.99 untuk *real*. *F1-score* yang tinggi, sekitar 0.98, menunjukkan bahwa model memiliki performa yang konsisten baik pada kedua kelas.

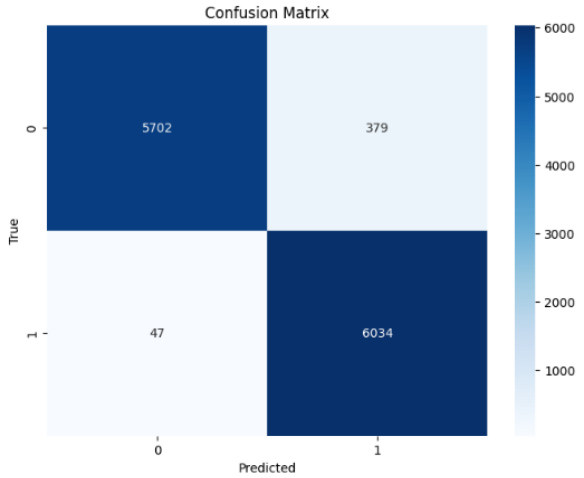



***Gambar 10.*** *Confusion matrix model EfficientNetB0 pada data validasi*

***Tabel 8.*** *Metrik evaluasi model EfficientNetB0 pada data validasi*

| Kelas | *Precision* | *Recall* | *F1-score* |
|---|---|---|---|
| *Fake* | 0.99 | 0.93 | 0.96 |
| *Real* | 0.94 | 0.99 | 0.96 |
| Akurasi | 0.96 | | |

Gambar 10 menunjukkan confusion matrix pada data validasi untuk model EfficientNetB0. Model ini berhasil mengklasifikasikan 5673 gambar *fake* dengan benar sebagai *true positive (TP)*, namun terdapat 407 *false negative (FN)* pada kategori *fake* dan 49 *false positive* (FP) pada kategori *real*. Sebagian besar kesalahan klasifikasi terjadi pada gambar *fake*, yang sering kali diklasifikasikan sebagai *real*. Metrik evaluasi yang tercantum dalam Tabel 8 menunjukkan *precision* sebesar 0.99 dan *recall* sebesar 0.93 untuk kategori *fake*, serta *precision* 0.94 dan *recall* 0.99 untuk kategori *real*. *F1-score* untuk kedua kelas mencapai 0.96, yang menunjukkan bahwa model memiliki kinerja yang cukup baik, meskipun *recall* untuk kategori *fake* lebih rendah dibandingkan dengan model ViT yang mencapai 0.97 pada dataset validasi.




***Gambar 11.*** *Confusion matrix model EfficientNetB0 pada data pengujian*

***Tabel 9.*** *Metrik evaluasi model EfficientNetB0 pada data pengujian*

| Kelas | *Precision* | *Recall* | *F1-score* |
|---|---|---|---|
| *Fake* | 0.99 | 0.94 | 0.96 |
| *Real* | 0.94 | 0.99 | 0.97 |
| Akurasi | 0.96 | | |

Gambar 11 menunjukkan *confusion matrix* untuk dataset pengujian. Model EfficientNetB0 mengklasifikasikan 5702 gambar *fake* dengan benar, namun mengalami 379 *false negatives (FN)* untuk kategori *fake* dan 47 *false positives (FP)* pada kategori *real*. Hasil ini menunjukkan pola kesalahan yang mirip dengan data validasi, yaitu lebih banyak kesalahan pada kategori *fake*. Metrik evaluasi pada Tabel 9 menunjukkan bahwa precision untuk kategori *fake* adalah 0.99 dan *recall* 0.94, sementara untuk kategori *real*, *precision* 0.94 dan *recall* 0.99. *F1-score* untuk kategori *real* sedikit lebih tinggi (0.97) dibandingkan dengan kategori *fake* (0.96), yang menunjukkan bahwa model ini sedikit lebih unggul dalam mengenali gambar asli dibandingkan dengan gambar palsu.

Jika dibandingkan dengan model ViT yang menunjukkan performa lebih baik dalam mengidentifikasi gambar *fake* dengan *precision* 0.99 dan *recall* 0.97, model EfficientNetB0 tampaknya lebih sering salah dalam mengidentifikasi gambar *fake*. Hal ini mungkin disebabkan oleh model ViT yang lebih mampu menangkap pola global dalam gambar, sementara EfficientNetB0 lebih fokus pada fitur lokal. Pada *dataset* pengujian, model ViT mengklasifikasikan dengan akurasi 98% dibandingkan dengan 96% pada model EfficientNetB0, menunjukkan bahwa meskipun kedua model memiliki keunggulan, ViT sedikit lebih baik dalam hal pengenalan gambar *fake* dan memiliki performa yang lebih konsisten.

Pada penelitian ini, kami membandingkan dua model, yaitu Vision Transformer (ViT) dan EfficientNetB0, yang dilatih menggunakan Google Colab dengan GPU T4. Kedua model memiliki perbedaan signifikan dalam hal efisiensi parameter, waktu pelatihan, serta ukuran file model yang diekspor.

***Tabel 10.*** *Perbandingan Jumlah parameter model Vision Transformer dan EfficientNetB0*

| Model | *Total Params* | *Trainable Params* | *Non-trainable Params* |
|---|---|---|---|
| *Vision Transformer* | 86,568,657 (330.23 MB) | 1,001 (3.91 KB) | 86,567,656 (330.23 MB) |
| *EfficientNetB0* | 4,131,620 (15.76 MB) | 4,089,597 (15.60 MB) | 42,023 (164.16 KB) |

Model ViT, seperti yang ditunjukkan dalam Tabel 10, memiliki total parameter sebesar 86.5 juta, dengan sebagian besar parameter tersebut tidak dapat dilatih (sekitar 330.23 MB), sementara hanya sekitar 1.001 parameter yang dapat dilatih, yang berukuran hanya 3.91 KB. Di sisi lain, model EfficientNetB0, yang lebih kecil dan lebih efisien dalam jumlah parameter, memiliki total parameter sekitar 4.1 juta dengan lebih dari 4 juta parameter dapat dilatih, yang menghasilkan ukuran model hanya 15.76 MB, seperti yang terlihat pada Tabel 10. Jumlah parameter yang jauh lebih kecil pada EfficientNetB0 menunjukkan bahwa model ini dirancang untuk efisiensi komputasi yang lebih baik, dengan mengorbankan sedikit kompleksitas dibandingkan dengan ViT.

***Tabel 11.*** *Perbandingan waktu training model Vision Transformer dan EfficientNetB0*

| | *Vision Transformer* | | *EfficienNetB0* | |
|---|---|---|---|---|
| *Epoch* | *Waktu (s)* | *Waktu per step (ms/step)* | *Waktu (s)* | *Waktu per step (ms/step)* |
| 1 | 640 | 266 | 303 | 97 |
| 2 | 622 | 273 | 209 | 91 |
| 3 | 623 | 273 | 211 | 93 |
| 4 | 625 | 274 | 261 | 92 |
| 5 | 624 | 274 | 209 | 92 |
| 6 | 623 | 273 | 212 | 93 |
| 7 | 625 | 274 | 211 | 92 |
| 8 | 624 | 274 | 261 | 92 |
| 9 | 623 | 273 | 211 | 92 |
| 10 | 624 | 274 | 264 | 93 |

Jika dibandingkan berdasarkan waktu pelatihan, model ViT memerlukan lebih banyak waktu untuk menyelesaikan satu epoch dibandingkan dengan EfficientNetB0. Tabel 11 menunjukkan bahwa setiap epoch pada ViT memakan waktu rata-rata sekitar 624 detik, dengan waktu per langkah mencapai 274 ms/step. Ini mengindikasikan bahwa ViT lebih kompleks dan memerlukan lebih banyak waktu untuk memproses data dalam setiap langkahnya. Sementara itu, model EfficientNetB0 yang ditampilkan pada Tabel 11 memiliki waktu pelatihan yang jauh lebih cepat, dengan rata-rata total waktu per epoch sekitar 211 detik dan waktu per langkah 92 ms/step. Kecepatan pelatihan yang lebih tinggi ini menunjukkan bahwa EfficientNetB0 lebih efisien dalam hal penggunaan waktu dan sumber daya komputasi.

***Tabel 12.*** *Perbandingan waktu prediksi data validasl model Vision Transformer dan EfficientNetB0*

| Model | *Waktu (s)* | *Waktu per step (ms/step)* |
|---|---|---|
| *Vision Transformer* | 146 | 165 |
| *EfficientNetB0* | 44 | 48 |

Selain itu, saat melakukan prediksi pada data validasi, model ViT membutuhkan 146 detik untuk memproses 6080 gambar, yang menghasilkan waktu per langkah sekitar 165 ms/step, sedangkan EfficientNetB0 memproses gambar yang sama hanya dalam 44 detik dengan waktu per langkah sekitar 48 ms/step. Hal ini menunjukkan bahwa EfficientNetB0 lebih cepat dalam hal inferensi dibandingkan dengan ViT, yang lebih memerlukan waktu untuk memberikan hasil prediksi.

Dalam hal ukuran model, ViT memiliki file yang jauh lebih besar setelah diekspor, mencapai 330,3 MB. Di sisi lain, model EfficientNetB0 memiliki ukuran yang jauh lebih kecil, hanya sekitar 47,9 MB. Ukuran model yang lebih kecil pada EfficientNetB0 memudahkan penyimpanan dan distribusi model, serta lebih menghemat ruang pada perangkat penyimpanan, menjadikannya pilihan yang lebih efisien untuk implementasi di perangkat dengan keterbatasan sumber daya.

Berdasarkan perbandingan ini, dapat disimpulkan bahwa meskipun model ViT menunjukkan performa yang sangat baik dalam hal akurasi dan metrik evaluasi lainnya, model EfficientNetB0 menawarkan efisiensi yang lebih baik dalam hal jumlah parameter, waktu pelatihan, dan ukuran file model yang diekspor.

Perbandingan hasil akurasi antara model Vision Transformer (ViT) dan EfficientNetB0 dengan penelitian terdahulu menunjukkan peningkatan performa signifikan. Model ViT dan EfficientNetB0 masing-masing mencapai akurasi sebesar 98% dan 96% pada dataset validasi, yang lebih unggul dibandingkan penelitian Mu et al. [5], yang menggunakan CNN dengan 6 lapisan convolutional dan hanya mencapai akurasi 91% dalam mendeteksi wajah asli dan deepfake.

Hasil akurasi ViT juga jauh melampaui arsitektur Transformer BEiT yang digunakan oleh Prawiratama et al. [6], yang hanya mencapai akurasi 80% pada klasifikasi gambar generatif dan gambar buatan tangan. Dibandingkan dengan pendekatan Bird dan Lotfi [7], yang mengembangkan CNN untuk mendeteksi gambar sintetis pada dataset CIFAKE dengan akurasi 92.98%, kedua model yang diimplementasikan dalam penelitian ini menunjukkan peningkatan yang signifikan.

Pendekatan Hybrid VGG16-CNN oleh Raza et al. [8], yang mencatatkan akurasi 94% untuk mendeteksi konten deepfake, juga masih berada di bawah kinerja ViT. Dengan akurasi yang lebih tinggi, model ViT dan EfficientNetB0 membuktikan keunggulannya dalam menangani klasifikasi antara gambar asli dan hasil generatif.

## 5. Penutup

Penelitian ini bertujuan untuk mengembangkan model klasifikasi berbasis EfficientNet-B0 dan Vision Transformer (ViT) untuk membedakan gambar asli dan gambar yang dihasilkan oleh kecerdasan buatan generatif. Hasil penelitian menunjukkan bahwa kedua model memiliki akurasi yang tinggi dalam mengklasifikasikan gambar, dengan ViT mencapai akurasi 98% dan EfficientNet-B0 mencapai akurasi 96%. Berdasarkan evaluasi metrik, model ViT menunjukkan performa lebih unggul dengan precision 99%, recall 97%, dan F1-score 98% untuk kategori fake, serta precision 97%, recall 99%, dan F1-score 98% untuk kategori real. Sementara itu, model EfficientNet-B0 memperoleh precision 99%, recall 93%, dan F1-score 96% untuk kategori fake, serta precision 94%, recall 99%, dan F1-score 96% untuk kategori real. Implikasi praktis dari penelitian ini adalah pengembangan teknologi yang lebih canggih untuk mendeteksi dan mengklasifikasikan gambar generatif, yang dapat digunakan untuk mengatasi tantangan era digital terkait manipulasi media visual. Namun, penelitian ini memiliki keterbatasan, seperti penggunaan dataset yang terbatas dan kebutuhan akan pengujian lebih lanjut pada berbagai jenis gambar generatif. Oleh karena itu, saran untuk penelitian selanjutnya adalah memperluas dataset yang digunakan dan menguji model pada berbagai jenis gambar generatif untuk meningkatkan generalisasi dan akurasi model. Dengan demikian, penelitian ini memberikan kontribusi signifikan dalam mengatasi tantangan manipulasi media visual di era digital.

## Referensi

[1] Keith McAleer, “AI is the New Electricity’: Insights from Dr. Andrew Ng - UC Berkeley Sutardja Center,” Berkeley SCET . Accessed: Mar. 21, 2025. [Online]. Available: https://scet.berkeley.edu/ai-is-the-new-electricity-insights-from-dr-andrew-ng/

[2] I. R. Dewi, "Foto Mesum Taylor Swift Buatan AI Viral, Begini Kronologinya," CNBC Indonesia. Accessed: Feb. 29, 2024. [Online]. Available: https://www.cnbcindonesia.com/tech/20240129113341-37-509781/foto-mesum-taylor-swift-buatan-ai-viral-begini-kronologinya

[3] A. Clark, "A real photo took two honors in an AI competition. Here's the inside story. - CBS News," CBS NEWS. Accessed: Sep. 22, 2025. [Online]. Available: https://www.cbsnews.com/news/real-photo-ai-competition-flamingone-miles-astray/

[4] A. Pocol, L. Istead, S. Siu, S. Mokhtari, and S. Kodeiri, "Seeing is No Longer Believing: A Survey on the State of Deepfakes, AI-Generated Humans, and Other Nonveridical Media," 2024, pp. 427–440. doi: 10.1007/978-3-031-50072-5_34.

[5] J. Mu, M. Adrezo, and A. N. Haikal, "Identifikasi Wajah Asli dan Buatan Deepfake Menggunakan Metode Convolutional Neural Network," *Teknika*, vol. 13, no. 1, pp. 45–50, Jan. 2024, doi: 10.34148/teknika.v13i1.705.

[6] R. A. Prawiratama, S. Sumarno, and I. A. Kautsar, "RANCANG BANGUN APLIKASI UJI KEMIRIPAN GAMBAR AI GENERATIVE DAN GAMBAR BUATAN TANGAN MENGGUNAKAN METODE DEEP LEARNING," *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, vol. 7, no. 1, p. 114, Jun. 2024, doi: 10.37600/tekinkom.v7i1.1192.

[7] 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.

[8] A. Raza, K. Munir, and M. Almutairi, "A Novel Deep Learning Approach for Deepfake Image Detection," *Applied Sciences*, vol. 12, no. 19, p. 9820, Sep. 2022, doi: 10.3390/app12199820.

[9] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *ArXiv*, 2019.

[10] D. Putri Ayuni, Jasril, M. Irsyad, F. Yanto, and S. Sanjaya, "AUGMENTASI DATA PADA IMPLEMENTASI CONVOLUTIONAL NEURAL NETWORK ARSITEKTUR EFFICIENTNET-B3 UNTUK KLASIFIKASI PENYAKIT DAUN PADI," *ZONAsi: Jurnal Sistem Informasi*, vol. 5, no. 2, pp. 239–249, May 2023, doi: 10.31849/zn.v5i2.13874.

[11] S. Aras, A. Setyanto, and Rismayani, "Deep Learning Untuk Klasifikasi Motif Batik Papua Menggunakan EfficientNet dan Trasnfer Learning," *Insect (Informatics and Security): Jurnal Teknik Informatika*, vol. 8, no. 1, pp. 11–20, Oct. 2022, doi: 10.33506/insect.v8i1.1865.

[12] I. Rizka Fadhillah, M. Muharrom Al Haromainy, and H. Maulana, "IMPLEMENTASI MODEL TRANSFER LEARNING EFFICIENTNET UNTUK PENDETEKSIAN BAHASA ISYARAT INDONESIA (BISINDO) PADA PERANGKAT ANDROID," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 4, pp. 7816–7822, Aug. 2024, doi: 10.36040/jati.v8i4.10463.

[13] D. Adam and H. Santoso, "IMAGE CLASSIFICATION OF HOUSEHOLD BENEFICIARIES OF DIRECT CASH ASSISTANCE USING EFFICIENTNET IN DKI JAKARTA PROVINCE," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp. 665–671, Aug. 2024, doi: 10.52436/1.jutif.2024.5.4.2121.

[14] A. N. Fajrina, Z. H. Pradana, S. I. Purnama, and S. Romadhona, "Penerapan Arsitektur EfficientNet-B0 Pada Klasifikasi Leukimia Tipe Acute Lymphoblastik Leukimia," *Jurnal Riset Rekayasa Elektro*, vol. 6, no. 1, p. 59, Jun. 2024, doi: 10.30595/jrre.v6i1.22090.

[15] W. R. PERDANI, R. MAGDALENA, and N. K. CAECAR PRATIWI, "Deep Learning untuk Klasifikasi Glaukoma dengan menggunakan Arsitektur EfficientNet," *ELKOMIKA: Jurnal*

*Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, vol. 10, no. 2, p. 322, Apr. 2022, doi: 10.26760/elkomika.v10i2.322.

[16] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[17] R. Uthama, Yuhandri, and Billy Hendrik, "Vision Transformer untuk Identifikasi 15 Variasi Citra Ikan Koi," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 5, no. 1, pp. 159–168, May 2024, doi: 10.37859/coscitech.v5i1.6711.

[18] A. Pangestu, B. Purnama, and R. Risnandar, "Vision Transformer untuk Klasifikasi Kematangan Pisang," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 75–84, Feb. 2024, doi: 10.25126/jtiik.20241117389.

[19] T. Febriyanto and S. Syofian, "Implementasi Deep Learning Menggunakan Vision Transformer Untuk Klasifikasi Penyakit Daun Padi," *Journal TIFDA (Technology Information and Data Analytic)*, vol. 1, no. 2, pp. 34–39, Dec. 2024, doi: 10.70491/tifda.v1i2.47.

[20] J. A. Figo, N. Yudistira, and A. W. Widodo, "Deteksi Covid-19 dari Citra X-ray menggunakan Vision Transformer," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 3, 2023.

[21] I. C. Sukandar, F. Tri Anggraeny, and M. Hanindia Prami Swari, "Effect of Optimisation in Brain Tumour Classification with CNN-VIT Hybrid," *Antivirus : Jurnal Ilmiah Teknik Informatika*, vol. 18, no. 1, pp. 112–124, Jun. 2024, doi: 10.35457/antivirus.v18i1.3557.

[22] COCO, "COCO - Common Objects in Context," COCO. Accessed: Nov. 24, 2024. [Online]. Available: https://cocodataset.org/#home

[23] J. Heldt, "SyntheticEye AI-Generated Images Dataset," Kaggle. Accessed: Nov. 05, 2024. [Online]. Available: https://www.kaggle.com/datasets/jacobheldt/syntheticeue-ai-generated-images-dataset

[24] A. A. Handoko, M. A. Rosid, and U. Indahyanti, "Implementasi Convolutional Neural Network (CNN) Untuk Pengenalan Tulisan Tangan Aksara Bima," *SMATIKA JURNAL*, vol. 14, no. 01, pp. 96–110, Jul. 2024, doi: 10.32664/smatika.v14i01.1196.

[25] EVIDENTY AI, "How to interpret a confusion matrix for a machine learning model," EVIDENTY AI. Accessed: Nov. 16, 2024. [Online]. Available: https://www.evidentlyai.com/classification-metrics/confusion-matrix