Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ 475-490

ANALISIS PERBANDINGAN CNN DAN VISION TRANSFORMER DALAM KLASIFIKASI HURUF *SEMAPHORE* PRAMUKA BERBASIS CITRA

Muh. Akbar Haeruddin[1], Desi Anggreani[2], Muhyiddin A.M Hayat[3]
[1,2,3]Universitas Muhammadiyah Makassar, Indonesia
[1]105841104622@student.unismuh.ac.id [2]desianggreani@unismuh.ac.id
[3]muhyiddin@unismuh.ac.id
*Coresponding Author

Info Artikel

Pengajuan : 31-07-2026
Peninjauan : 10-08-2026
Revisi : 26-08-2026
Diterima : 30-08-2026
Diterbitkan : 20-09-2026

Kata Kunci:
Convolutional Neural Network, DeiT-Tiny, EfficientNet-B0, Klasifikasi Citra, Semaphore Pramuka, Vision Transformer.

Abstrak

Klasifikasi huruf Semaphore Pramuka merupakan salah satu permasalahan computer vision yang menantang karena kemiripan pose antarkelas, variasi subjek, pencahayaan, serta latar belakang citra. Penelitian ini bertujuan membandingkan kinerja Convolutional Neural Network (CNN) menggunakan model EfficientNetB0 dan Vision Transformer (ViT) menggunakan model DeiT-Tiny dalam mengklasifikasikan citra huruf Semaphore. Dataset terdiri atas 26 kelas huruf A–Z dengan 1.014 citra untuk pelatihan dan validasi serta 234 citra pengujian eksternal. Tahap praproses meliputi konversi citra ke format RGB, resize menjadi 224×224 piksel, normalisasi, dan augmentasi pada data pelatihan. Kedua model dilatih menggunakan transfer learning dengan konfigurasi pelatihan yang setara. Proses pelatihan menggunakan optimizer AdamW dengan learning rate 3×10^{-4}, weight decay 1×10^{-4}, batch size 64, penjadwalan cosine annealing, serta label smoothing 0,1, dipadukan dengan early stopping untuk mencegah overfitting. Kedua model kemudian dievaluasi menggunakan metrik accuracy, precision, recall, F1-score, dan confusion matrix. Hasil pengujian menunjukkan bahwa DeiT-Tiny memberikan performa yang lebih baik dibandingkan EfficientNet-B0. DeiT-Tiny memperoleh accuracy 97,01%, precision 97,64%, recall 97,01%, dan F1-score 96,95%, sedangkan EfficientNetB0 memperoleh accuracy 82,05%, precision 85,48%, recall 82,05%, dan F1-score 81,53%. Selain itu, DeiT-Tiny mampu mengklasifikasikan 23 dari 26 kelas secara sempurna dengan jumlah kesalahan yang lebih sedikit. Hasil penelitian menunjukkan bahwa Vision Transformer berbasis DeiT-Tiny lebih efektif dibandingkan EfficientNetB0 dalam klasifikasi citra huruf Semaphore pada dataset yang digunakan. Temuan ini secara teoretis memperkuat bukti bahwa mekanisme self-attention pada ViT mampu menangkap relasi spasial global antarbagian pose secara lebih baik pada dataset berskala kecil, sekaligus memberikan kontribusi praktis sebagai dasar pengembangan sistem pengenalan Semaphore berbasis citra yang lebih akurat, misalnya untuk media pembelajaran maupun penilaian keterampilan Semaphore secara otomatis

How to Cite : Haeruddin, M. A., Anggreani, D., & Hayat, M. A. M. (2026). Analisis Perbandingan CNN Dan Vision Transformer Dalam Klasifikasi Huruf Semaphore Pramuka Berbasis Citra. Journal of Computer Science and Information Technology (JCSIT), 3(4), 475–490. https://doi.org/10.70248/jcsit.v3i4.4366

DOI : https://doi.org/10.70248/jcsit.v3i4.4366
This is an open access article under https://creativecommons.org/licenses/by/4.0/
Copyright© 2026 by Author. Published by Yayasan Nuraini Ibrahim Mandiri

PENDAHULUAN

Komunikasi visual melalui gerakan tubuh dan tangan berperan penting dalam berbagai bidang, termasuk pendidikan, militer, dan organisasi kepemudaan. [1] menyatakan bahwa komunikasi nonverbal membantu penyampaian pesan dalam interaksi lintas budaya dan profesi. Perkembangan *computer vision* dan *deep learning* kemudian membuka peluang untuk mengenali gestur secara otomatis melalui citra digital.

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ 475-490

Salah satu bentuk komunikasi visual tersebut adalah *semaphore* Pramuka, yaitu sistem yang merepresentasikan huruf A sampai Z melalui posisi dan orientasi kedua lengan yang memegang bendera. Sistem ini tidak bergantung pada perangkat elektronik sehingga tetap dapat digunakan dalam komunikasi lapangan dan kondisi darurat [2]. Namun, perbedaan antarkelas *semaphore* sering kali hanya terletak pada sudut dan arah lengan. Kemiripan pose, variasi subjek, pencahayaan, latar belakang, dan sudut pengambilan gambar menjadikan klasifikasi huruf *semaphore* sebagai permasalahan yang menantang [3].

Convolutional Neural Network (CNN) banyak digunakan untuk pengenalan gestur karena mampu mempelajari fitur lokal citra secara hierarkis. [4] menerapkan *MS-MobileNet* dengan *multi-scale convolution* dan *depth-separable convolution* serta memperoleh akurasi antara 88,41% dan 98,26% pada beberapa dataset gestur. [5] juga menunjukkan bahwa arsitektur CNN, seperti LeNet, AlexNet, VGG, Inception, dan *ResNet*, telah digunakan secara luas dalam klasifikasi citra. Meskipun demikian, CNN masih dapat mengalami sensitivitas terhadap perubahan orientasi dan kesulitan dalam memodelkan hubungan spasial global, terutama pada data terbatas dan kelas dengan perbedaan visual yang kecil [6].

Penelitian yang secara khusus membahas *semaphore* telah dilakukan oleh [7] menggunakan segmentasi warna HSV, ekstraksi fitur *Discrete Wavelet Transform* 2D, dan *Back Propagation Neural Network*. Metode tersebut menghasilkan akurasi 94% pada jarak 3 meter, tetapi menurun menjadi 83% pada jarak 7 meter. Hasil tersebut membuktikan bahwa pengenalan *semaphore* dapat dilakukan secara otomatis, tetapi pendekatannya masih bergantung pada ekstraksi fitur manual dan belum membandingkan arsitektur *deep learning* modern.

Sebagai alternatif CNN, *Vision Transformer* (ViT) memproses citra menjadi sejumlah *patch* dan menggunakan mekanisme *self-attention* untuk mempelajari hubungan global antarbagian citra. Akan tetapi, ViT umumnya membutuhkan data pelatihan dalam jumlah besar karena memiliki *inductive bias* yang lebih rendah daripada CNN [8]. [9] menunjukkan bahwa integrasi *depth-wise convolution* pada ViT-Tiny dapat meningkatkan akurasi sebesar 2,4% pada CIFAR-10 dan mengungguli model yang berparameter lebih besar. [10] juga menunjukkan potensi arsitektur *transformer* dalam klasifikasi gestur tangan. Namun, penelitian tersebut menggunakan dataset umum dan belum berfokus pada huruf *semaphore* Pramuka dengan kemiripan visual antarkelas yang tinggi.

Selain ketepatan klasifikasi, efisiensi komputasi merupakan pertimbangan penting dalam pemilihan model pengenalan gestur lapangan. Sistem *semaphore* berpotensi diterapkan pada perangkat cerdas atau perangkat edge yang memiliki keterbatasan daya, memori, dan kemampuan komputasi. Jumlah parameter dan ukuran bobot menentukan kebutuhan penyimpanan serta memori, sedangkan FLOPs menggambarkan beban operasi untuk satu kali inferensi. Namun, FLOPs tidak selalu berbanding lurus dengan waktu eksekusi karena latency juga dipengaruhi oleh paralelisme, bandwidth memori, implementasi operator, dan perangkat keras. Oleh karena itu, penelitian ini tidak hanya membandingkan kinerja klasifikasi *EfficientNetB0* dan *DeiT-Tiny*, tetapi juga menganalisis parameter total, ukuran bobot, FLOPs, dan waktu inferensi per citra. Analisis tersebut diperlukan untuk menilai kompromi antara akurasi dan efisiensi sebelum model diterapkan pada sistem pengenalan *semaphore* secara real-time.

Berdasarkan penelitian terdahulu, masih diperlukan kajian yang membandingkan CNN dan ViT secara langsung pada dataset *semaphore* Pramuka yang

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ 475-490

dikumpulkan secara mandiri. Oleh karena itu, penelitian ini membandingkan *EfficientNetB0* sebagai representasi CNN dan *DeiT-Tiny* sebagai representasi ViT melalui *transfer learning* dan anggaran penyetelan hiperparameter yang setara. Pembagian data pelatihan dan validasi dilakukan berdasarkan subjek untuk mencegah kebocoran data, sedangkan evaluasi akhir menggunakan data uji terpisah. Kinerja kedua model dianalisis menggunakan akurasi, presisi berbobot, *recall* berbobot, *F1-score* berbobot, dan *confusion matrix*. Hasil penelitian diharapkan memberikan bukti empiris mengenai arsitektur yang lebih sesuai untuk klasifikasi huruf *semaphore* serta menjadi referensi bagi pengembangan sistem pengenalan *semaphore* berbasis citra.

Tabel 1. Penelitian Terkait dan *Research Positioning*

Referensi	Fokus	Pendekatan	Keterbatasan
Niu et al. (2025)	Pengenalan gesture tangan	*CNN* (MSMobileNet)	Belum membahas *Semaphore* dan *ViT*.
Referensi	Fokus	Pendekatan	Keterbatasan
Putra et al. (2024)	Pengenalan huruf *Semaphore*	DWT + BPNN	Belum menggunakan *deep learning modern*.
Tang (2023)	Tinjauan klasifikasi citra	Review *CNN* dan GCN	Tidak melakukan implementasi model.
Zhang et al. (2025)	*Vision Transformer*	DWConv + *ViT*	Tidak membahas klasifikasi *Semaphore*.
Penelitian ini	Klasifikasi huruf *Semaphore*	*EfficientNetB0* dan *DeiT-Tiny*	Membandingkan *CNN* dan *Vision Transformer*.

Berdasarkan tinjauan pada Tabel 1, terlihat adanya kesenjangan penelitian (*research gap*) pada bidang klasifikasi citra huruf *Semaphore*. Penelitian oleh Niu (2025) berfokus pada pengenalan *gesture* tangan menggunakan arsitektur CNN berbasis *MS-MobileNet*, namun belum membahas klasifikasi huruf *Semaphore* maupun penerapan *Vision Transformer*. Putra et al. (2024) telah mengembangkan sistem pengenalan huruf *Semaphore* menggunakan kombinasi *Discrete Wavelet Transform* (DWT) dan *Back Propagation Neural Network* (BPNN), tetapi masih menggunakan metode ekstraksi fitur dan klasifikasi konvensional. Sementara itu, Tang (2023) menyajikan tinjauan mengenai perkembangan berbagai arsitektur CNN dan GCN tanpa melakukan implementasi maupun evaluasi eksperimental. Di sisi lain, Zhang et al. (2025) mengusulkan peningkatan performa *Vision Transformer* melalui integrasi *Depth-Wise Convolution*, namun penelitian tersebut tidak diterapkan pada kasus klasifikasi huruf *Semaphore* maupun dibandingkan secara langsung dengan model CNN.

Berdasarkan kesenjangan penelitian tersebut, penelitian ini mengimplementasikan dan membandingkan dua arsitektur *deep learning*, yaitu *EfficientNetB0* sebagai representasi Convolutional Neural Network (CNN) dan *DeiT-Tiny* sebagai representasi Vision Transformer (ViT) pada klasifikasi citra huruf *Semaphore*. Perbandingan dilakukan menggunakan dataset yang sama, konfigurasi pelatihan yang setara, serta dievaluasi berdasarkan metrik accuracy, precision, recall, F1-score, dan confusion matrix, sehingga diperoleh analisis yang objektif mengenai performa kedua arsitektur dalam menyelesaikan tugas klasifikasi huruf *Semaphore*.

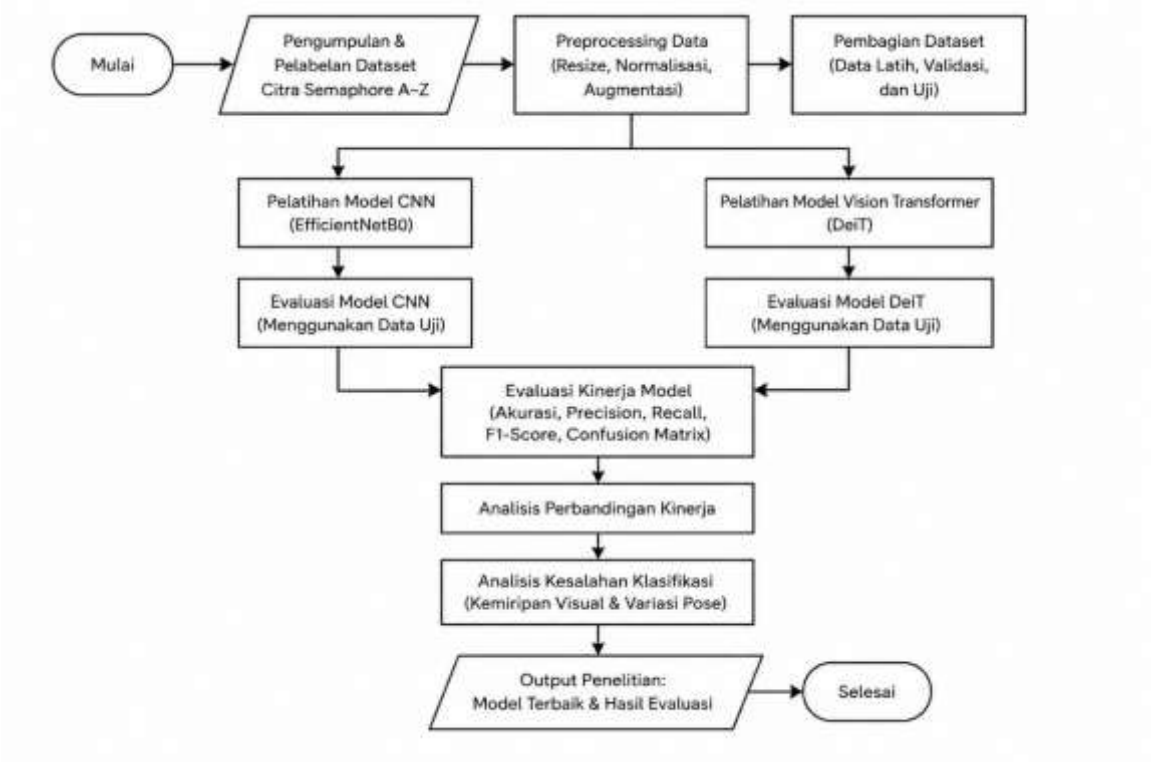
Berdasarkan kesenjangan tersebut, penelitian ini dirumuskan ke dalam

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 475-490

beberapa pertanyaan penelitian (*research questions*) berikut:

1. Bagaimana proses implementasi Convolutional Neural Network (CNN) dan Vision Transformer (ViT) dalam klasifikasi citra *Semaphore* Pramuka menggunakan dataset yang disusun secara mandiri?
2. Bagaimana kemampuan Convolutional Neural Network (CNN) dan Vision Transformer (ViT) dalam mengklasifikasikan citra *Semaphore* Pramuka berdasarkan hasil evaluasi kinerja model?

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen untuk membandingkan kinerja Convolutional Neural Network (CNN) *EfficientNetB0* dan Vision Transformer *DeiT-Tiny* dalam mengklasifikasikan citra huruf *semaphore* Pramuka. Tahapan penelitian meliputi pengumpulan dan pemeriksaan dataset, praproses citra, pembagian data berdasarkan subjek, penyetelan hiperparameter, pelatihan model, serta evaluasi menggunakan data pengujian eksternal. Dataset terdiri atas 26 kelas huruf A–Z. Dataset pelatihan dan validasi berjumlah 1.014 citra dari 13 subjek, sedangkan dataset pengujian terdiri atas 234 citra terpisah. Performa model dievaluasi menggunakan accuracy, weighted precision, weighted recall, weighted F1-score, dan confusion matrix.



Gambar 1. Alur penelitian klasifikasi citra huruf *semaphore* Pramuka

Berdasarkan Gambar 1, penelitian dimulai dengan memasukkan dataset citra *semaphore* yang telah dikelompokkan ke dalam 26 folder kelas, yaitu huruf A sampai Z. Konsistensi label diperiksa melalui kesesuaian nama folder dan nama berkas. Program juga memeriksa kelengkapan data setiap subjek serta kemungkinan citra identik pada subjek atau kelas yang berbeda.

Pada tahap praproses, seluruh citra diubah menjadi format RGB dan diseragamkan ukurannya menjadi 224×224 piksel. Nilai piksel kemudian dinormalisasi menggunakan nilai mean (0,485; 0,456; 0,406) dan standard deviation (0,229; 0,224; 0,225) ImageNet. Augmentasi hanya diterapkan pada data pelatihan, meliputi translasi, perubahan skala, color jitter, Gaussian blur, dan penambahan Gaussian noise. Horizontal flip tidak digunakan karena pembalikan

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 475-490

arah lengan berpotensi mengubah makna pose *semaphore*.

Pembagian data pelatihan dan validasi dilakukan berdasarkan identitas subjek menggunakan pendekatan berbasis grup dengan random seed 42. Seluruh citra milik satu subjek hanya ditempatkan pada satu bagian data untuk mencegah kebocoran subjek (subject leakage).

Agar perbandingan bersifat adil (*fair comparison*), kedua arsitektur dilatih menggunakan skema *transfer learning* dari bobot pra-latih ImageNet dengan konfigurasi hiperparameter yang identik. Rincian lengkap hiperparameter pelatihan disajikan pada Tabel 2 untuk memastikan prinsip keterulangan eksperimen (*reproducibility*).

Tabel 2. Spesifikasi Hiperparameter Pelatihan

Komponen	Spesifikasi
Arsitektur CNN	*EfficientNetB0* (pra-latih ImageNet)
Arsitektur ViT	*DeiT-Tiny* / deit_tiny_patch16_224 (pra-latih ImageNet)
Strategi pelatihan	*Transfer learning* (*fine-tuning* seluruh lapisan)
Komponen	Spesifikasi
Ukuran input citra	224 × 224 × 3 (RGB)
Optimizer	AdamW
Learning rate awal	3×10^{-4} (0,0003)
Weight decay	1×10^{-4}
Skema penjadwalan LR (*decay*)	*Cosine Annealing* (CosineAnnealingLR)
Loss function	*Cross-Entropy Loss* dengan *label smoothing* = 0,1
Batch size	64
Jumlah epoch (maksimum)	30
Early stopping	*patience* = 8 (dipantau dari akurasi validasi)
Mixed precision	BF16 (*Automatic Mixed Precision*)
Random seed	42

Dengan konfigurasi yang seragam tersebut, perbedaan kinerja yang teramati dapat diatribusikan pada perbedaan arsitektur (CNN vs ViT), bukan pada perbedaan pengaturan pelatihan.

HASIL PENELITIAN DAN PEMBAHASAN

Pengumpulan Data

Dataset penelitian diperoleh melalui pengumpulan citra pose *semaphore* Pramuka yang merepresentasikan 26 huruf alfabet, yaitu A sampai Z. Seluruh citra dikelompokkan ke dalam folder berdasarkan label hurufnya sehingga proses pembacaan label dapat dilakukan secara otomatis. Sebelum digunakan, *Dataset* diperiksa untuk memastikan kesesuaian nama berkas dengan label folder, kelengkapan jumlah citra setiap subjek, serta tidak adanya citra identik yang tersebar pada kelas atau subjek berbeda.

Dataset pelatihan dan validasi terdiri atas 1.014 citra dari 13 subjek. Setiap subjek menyumbangkan tiga citra untuk masing-masing kelas sehingga terdapat 39 citra per kelas. Selain itu, disediakan 234 citra pengujian eksternal dengan sembilan citra pada setiap kelas. Distribusi yang seimbang membuat setiap kelas

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 475-490

mempunyai kontribusi yang sama terhadap proses pelatihan dan evaluasi.

Tabel 2. Distribusi *Dataset* Peneliti

Data	Subjek	Citra per kelas	Jumlah
Pelatihan	10	30	780
Validasi	3	9	234
Pengujian eksternal	Terpisah	9	234

Pembagian pelatihan dan validasi dilakukan berdasarkan identitas subjek. Sebanyak 10 subjek ditempatkan pada data pelatihan dan tiga subjek pada data validasi. Seluruh citra dari satu subjek hanya ditempatkan pada satu bagian data. Strategi ini digunakan agar model tidak dievaluasi menggunakan citra lain dari orang yang telah dilihat selama pelatihan. Data pengujian eksternal tidak digunakan dalam proses *Tuning*, pemilihan konfigurasi, maupun penentuan checkpoint.

Visualisasi sampel citra menunjukkan bahwa setiap huruf memiliki pola arah lengan dan posisi bendera yang berbeda. Namun, beberapa kelas hanya dibedakan oleh perubahan sudut lengan yang relatif kecil. Selain itu, terdapat variasi latar belakang, pencahayaan, jarak pengambilan gambar, postur tubuh, dan posisi subjek. Kondisi tersebut membuat model harus mempelajari pola pose utama sekaligus mengabaikan variasi visual yang tidak berkaitan langsung dengan kelas.

Visualisasi sampel memperlihatkan bahwa setiap kelas dibedakan terutama oleh arah kedua lengan dan posisi bendera. Variasi subjek, latar belakang, pencahayaan, dan proporsi tubuh menambah keragaman intrakelas. Sebaliknya, beberapa huruf mempunyai perbedaan sudut lengan yang kecil sehingga meningkatkan kemungkinan kesalahan antarkelas.

Tabel 3. Sample Huruf *Semaphore*

Huruf	Gambar	Huruf	Gambar
A		B	
Huruf	Gambar	Huruf	Gambar
C		D	

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 475-490

Praproses Data dan Augmentasi

Tahap praproses diawali dengan konversi citra menjadi format RGB dan penyesuaian ukuran menjadi 224 × 224 piksel. Ukuran tersebut digunakan karena sesuai dengan ukuran masukan standar *EfficientNetB0* dan *DeiT-Tiny*. Penyeragaman ukuran juga memastikan seluruh citra dapat diproses dalam satu *batch* tanpa perbedaan dimensi.

Augmentasi hanya diterapkan pada data pelatihan. Transformasi yang digunakan mencakup random *resized* crop dengan skala 0,95–1,05, color *jitter*, *Gaussian blur, dan random erasing*. Rotasi dan *horizontal flip* tidak digunakan karena arah pose merupakan informasi semantik pada semaphore. Membalik atau memutar citra berpotensi mengubah pose menjadi representasi huruf lain dan menghasilkan label yang tidak sesuai.

Tabel 4 menunjukkan contoh perubahan ukuran citra sebelum dan sesudah tahap resize menjadi 224 × 224 piksel. Proses *resize* dilakukan untuk menyamakan dimensi seluruh citra agar sesuai dengan ukuran masukan model *EfficientNetB0* dan *DeiT-Tiny.*

Tabel 4. Sample *resize Dataset*

Sebelum	Sesudah
	

Implementasi Perbandinga Metode *EfficientNetB0* dan *DeiT-Tiny*

Penelitian membandingkan *EfficientNetB0* sebagai representasi *CNN* dan *DeiT-Tiny* sebagai representasi *Vision Transformer*. *EfficientNetB0* mengekstraksi fitur secara hierarkis melalui operasi konvolusi dan memiliki 4.040.854 parameter setelah lapisan klasifikasi disesuaikan menjadi 26 keluaran. *DeiT-Tiny* membagi citra menjadi 196 patch berukuran 16 × 16 piksel, memprosesnya melalui 12 blok transformer encoder dengan tiga attention head, dan memiliki 5.529.434 parameter.

EfficientNetB0 dan *DeiT-Tiny* dipilih karena keduanya merupakan varian dasar yang relatif ringan dari dua keluarga arsitektur dengan mekanisme representasi berbeda. *EfficientNetB0* menggunakan compound scaling untuk menyeimbangkan kedalaman, lebar, dan resolusi jaringan, serta dalam implementasi penelitian ini memiliki 4.040.854 parameter. Model tersebut dipilih sebagai representasi CNN karena mampu mengekstraksi fitur lokal secara hierarkis dengan kebutuhan parameter yang relatif rendah. *DeiT-Tiny* memiliki 5.529.434 parameter dan dipilih sebagai representasi Vision Transformer ringan karena dirancang agar pelatihan transformer lebih efisien terhadap data dibandingkan ViT konvensional, terutama ketika menggunakan bobot pralatih dan transfer learning. Kedekatan skala parameter kedua model juga mengurangi

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 475-490

ketimpangan kapasitas dalam perbandingan CNN dan transformer.

MobileNetV3 tidak dipilih karena arsitektur tersebut terutama dioptimalkan melalui pencarian arsitektur dan operator yang berorientasi pada perangkat seluler, sedangkan fokus penelitian ini adalah membandingkan inductive bias konvolusi dengan self-attention dalam anggaran model yang relatif setara. Swin Transformer-Tiny juga tidak digunakan karena struktur hierarchical shifted-window dan kapasitasnya akan menambah faktor arsitektural serta beban komputasi yang berbeda dari *DeiT-Tiny*. Dengan demikian, pasangan *EfficientNetB0* dan *DeiT-Tiny* memberikan perbandingan yang lebih terkontrol antara pemrosesan fitur lokal berbasis konvolusi dan hubungan spasial global berbasis self-attention.

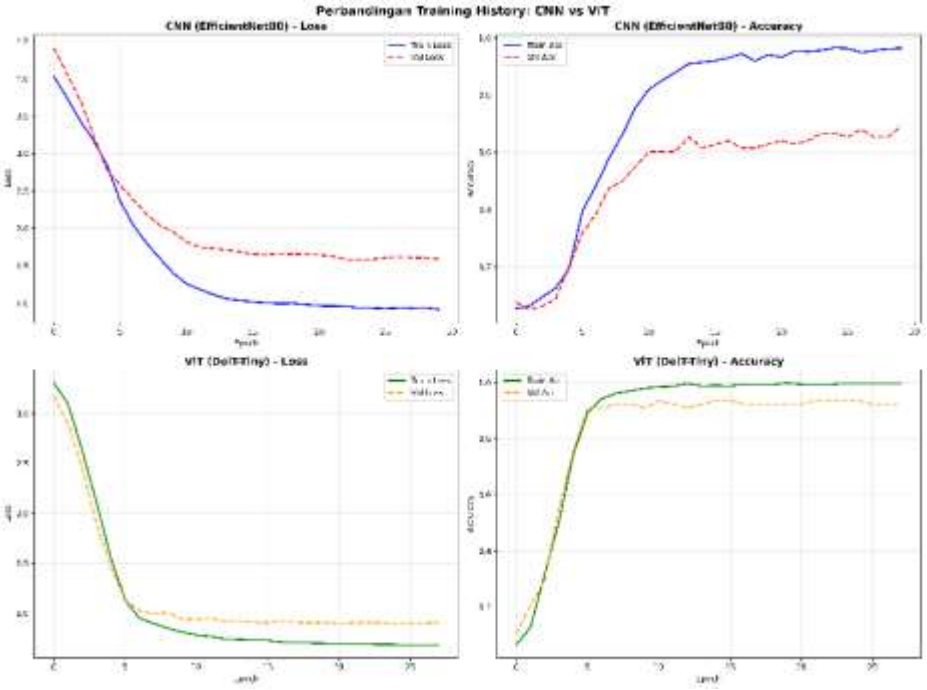
Kedua model menggunakan bobot pralatih serta menjalani protokol *Tuning* dengan anggaran yang sama. Tiga konfigurasi diuji pada tiga pembagian validasi berbasis subjek. Dengan demikian, masing-masing model menjalani sembilan percobaan. Konfigurasi dipilih berdasarkan rata-rata akurasi validasi tertinggi, kemudian rata-rata validation loss terendah digunakan sebagai pembanding apabila terdapat hasil yang sama.

Tabel 5. Hasil *Tuning EfficientNetB0* dan *DeiT-Tiny*

Model	Konfigurasi	Akurasi validasi	*Validation loss*
EfficientNetB0	T1	70,51% ± 6,54%	1,5256 ± 0,1354
EfficientNetB0	T2	32,05% ± 6,54%	2,4315 ± 0,1887
EfficientNetB0	T3	14,10% ± 6,54%	3,1609 ± 0,2727
DeiT-Tiny	T1	95,73% ± 2,63%	0,8321 ± 0,0606
DeiT-Tiny	T2	94,02% ± 0,60%	0,9100 ± 0,0453
DeiT-Tiny	T3	86,32% ± 4,83%	1,2150 ± 0,0282

Konfigurasi T1 memberikan rata-rata akurasi validasi tertinggi dan validation loss terendah pada kedua model. Oleh karena itu, konfigurasi tersebut digunakan untuk pelatihan akhir. Hasil ini menunjukkan bahwa pembekuan backbone dan regularisasi yang lebih kuat pada T2 serta T3 tidak memberikan peningkatan pada ruang konfigurasi yang diuji. Perbedaan hasil tersebut menunjukkan bahwa pembaruan parameter backbone masih memberikan kontribusi terhadap proses adaptasi model pada dataset penelitian ini. Pemilihan T1 selanjutnya digunakan sebagai dasar pelatihan akhir karena memberikan kinerja validasi yang lebih baik dibandingkan konfigurasi lainnya. Hasil tuning ini menjadi acuan dalam menentukan konfigurasi pelatihan yang digunakan pada tahap evaluasi menggunakan data testing.

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 475-490



Gambar 2. *Learning Curve*

EfficientNetB0 menyelesaikan 30 epoch. Validation loss terendah sebesar 1,5847 diperoleh pada epoch ke-23 dengan akurasi validasi 64,10%. Pada epoch tersebut, akurasi pelatihan mencapai 95,13%, sehingga terdapat selisih sebesar 31,03 poin persentase. Jarak yang cukup besar antara hasil pelatihan dan validasi menunjukkan indikasi bahwa *EfficientNetB0* lebih banyak menyesuaikan diri terhadap karakteristik data pelatihan.

Kurva memperlihatkan bahwa *DeiT-Tiny* menurunkan validation loss dengan lebih cepat. Akurasi validasinya telah melampaui 90% sejak epoch ke-6 dan kemudian relatif stabil pada kisaran 91,03–93,59%. Sebaliknya, akurasi validasi *EfficientNetB0* meningkat lebih lambat dan berfluktuasi pada kisaran yang lebih rendah. Temuan ini menjadi bukti empiris bahwa *DeiT-Tiny* memiliki konvergensi serta kestabilan validasi yang lebih baik dalam eksperimen ini.

Evaluasi akhir dilakukan terhadap 234 citra pengujian eksternal setelah konfigurasi dan checkpoint terbaik ditentukan. Precision, recall, dan F1-score dihitung menggunakan weighted average agar kontribusi setiap kelas disesuaikan dengan jumlah sampelnya. Selain itu, confusion matrix digunakan untuk melihat distribusi prediksi benar dan kesalahan klasifikasi pada masing-masing kelas. Hasil evaluasi tersebut selanjutnya digunakan untuk membandingkan kinerja *EfficientNetB0* dan *DeiT-Tiny* pada data yang belum digunakan selama proses pelatihan.

Tabel 6. Hasil Evaluasi Keseluruhan Model

Model	Akurasi	Precision	Recall	F1-score	Benar	Salah
EfficientNetB0	82,05%	85,48%	82,05%	81,53%	192	42
DeiT-Tiny	97,01%	97,64%	97,01%	96,95%	227	7

Efisiensi komputasi dianalisis menggunakan citra masukan 224 × 224 piksel dan batch size 1. Parameter total dihitung dari seluruh parameter model, ukuran bobot diperoleh dari ukuran aktual checkpoint, dan FLOPs dihitung untuk

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 475-490

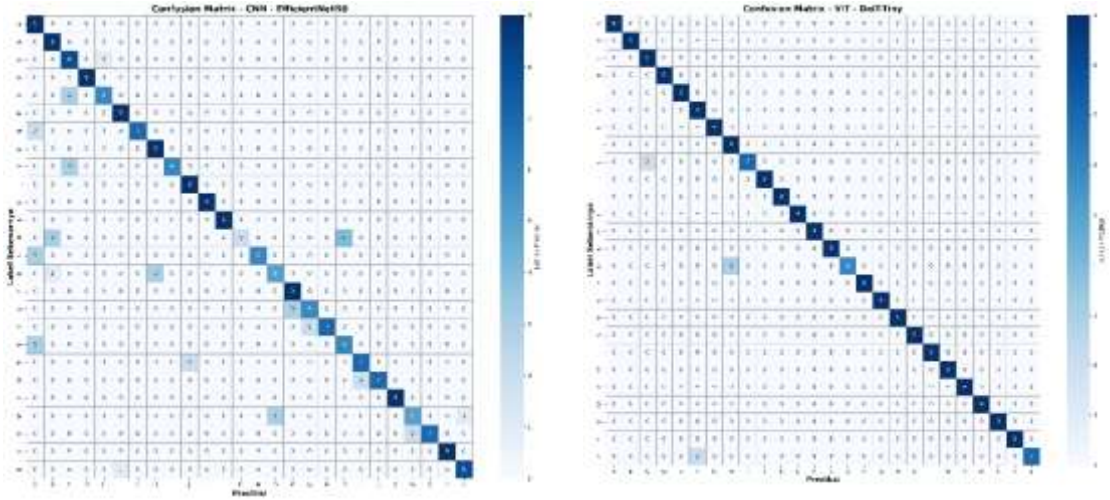
satu forward pass. Latency diukur pada GPU NVIDIA L40S menggunakan BF16 setelah tahap warm-up, dengan sinkronisasi CUDA pada setiap rangkaian pengukuran. Praproses dan waktu pemuatan berkas tidak dimasukkan sehingga nilai latency merepresentasikan komputasi model.

Tabel 7. Perbandingan Komputasi *EfficientNetB0* dan *DeiT-Tiny*

Model	Parameter	Ukuran bobot (MB)	FLOPs (GFLOPs)	Latency (ms/citra)	Perangkat
EfficientNet-B0	4.040.854	16,46	0,398	5,814 ± 0,064	NVIDIA L40S
DeiT-Tiny	5.529.434	22,18	1,258	4,655 ± 0,085	NVIDIA L40S

Berdasarkan Tabel 10, *DeiT-Tiny* memiliki jumlah parameter 36,84% lebih banyak, ukuran bobot 34,71% lebih besar, dan kebutuhan FLOPs sekitar 3,16 kali EfficientNet-B0. Meskipun demikian, *DeiT-Tiny* menghasilkan latency lebih rendah pada GPU NVIDIA L40S, yaitu 4,655 ms per citra dibandingkan 5,814 ms pada EfficientNet-B0. Hal ini menunjukkan bahwa FLOPs tidak selalu berbanding lurus dengan waktu inferensi aktual karena latency turut dipengaruhi oleh karakteristik operator, tingkat paralelisme, utilisasi Tensor Core, dan optimasi perangkat keras. Dengan demikian, *EfficientNetB0* lebih efisien dalam jumlah operasi dan kebutuhan penyimpanan, sedangkan *DeiT-Tiny* memberikan kombinasi akurasi dan latency GPU yang lebih baik pada perangkat pengujian. Namun, hasil latency pada GPU L40S tidak dapat langsung digeneralisasi ke perangkat edge karena perbedaan karakteristik perangkat keras.

EfficientNetB0 berhasil mengklasifikasikan 192 citra dan melakukan 42 kesalahan. *DeiT-Tiny* menghasilkan 227 prediksi benar dan hanya tujuh prediksi salah. *DeiT-Tiny* unggul sebesar 14,96 poin pada *accuracy*, 12,16 poin pada precision, 14,96 poin pada recall, dan 15,42 poin pada F1-score. Keunggulan pada seluruh metrik menunjukkan bahwa peningkatan tidak hanya terjadi pada jumlah prediksi benar, tetapi juga pada keseimbangan ketepatan dan sensiti*ViT*as klasifikasi antarkelas.



Gambar 3. *Confusion matrix EfficientNetB0* dan *DeiT-Tiny*

Pada *confusion matrix*, baris menunjukkan kelas aktual dan kolom menunjukkan kelas prediksi. Sebagian besar nilai berada pada diagonal utama, tetapi masih terdapat penyebaran kesalahan pada sejumlah kelas. Kelas M menjadi kelas tersulit karena hanya dua dari sembilan citra dikenali dengan benar. Empat citra M diprediksi sebagai S dan tiga citra diprediksi sebagai B.

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 475-490

Kelas O hanya menghasilkan lima prediksi benar, dengan tiga citra diprediksi sebagai H dan satu citra sebagai W. Kelas W juga hanya dikenali sebanyak lima citra, sedangkan tiga citra diprediksi sebagai O dan satu citra sebagai C. Pola ini menunjukkan bahwa *EfficientNetB0* masih kesulitan membedakan beberapa pose dengan konfigurasi arah lengan yang relatif mirip. Meskipun demikian, kelas D, K, L, dan V termasuk kelas yang dapat dikenali tanpa kesalahan dan tidak menerima prediksi salah dari kelas lain.

Confusion matrix DeiT-Tiny menunjukkan konsentrasi nilai yang lebih kuat pada diagonal utama. Sebanyak 23 dari 26 kelas dikenali secara sempurna. Kesalahan hanya terjadi pada kelas I, O, dan Z. Dua citra aktual I diprediksi sebagai C, tiga citra aktual O diprediksi sebagai H, dan dua citra aktual Z diprediksi sebagai F.

Kelas O menjadi kelas dengan recall terendah pada *DeiT-Tiny*, yaitu 66,67%, karena hanya enam dari sembilan citra yang dikenali dengan benar. Walaupun demikian, tidak terdapat citra dari kelas lain yang salah diprediksi sebagai O sehingga precision kelas O tetap tinggi. Pola tersebut memperlihatkan bahwa *DeiT-Tiny* lebih konsisten dalam membedakan sebagian besar pose *semaphore* dibandingkan *EfficientNetB0*.

Tabel 8. hasil setiap *EfiicientNetB0* kelas beserta nilai TP, TN, FP, dan FN

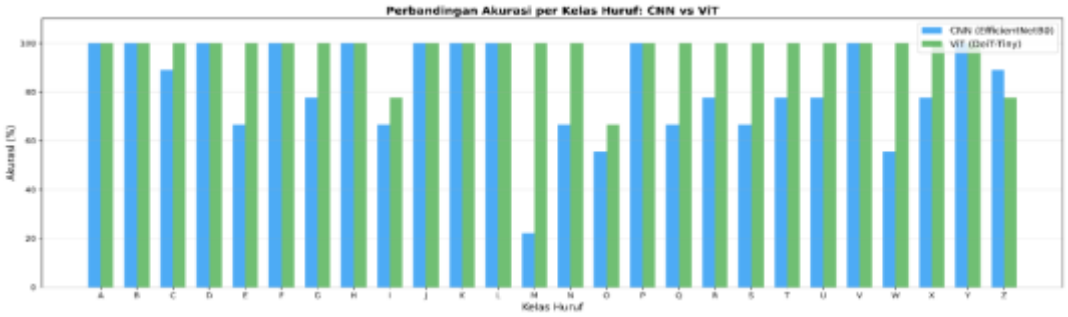
Kelas	Precision	Recall	F1-score	TP	TN	FP	FN	Sampel
D, K, L, V, Y	1,00	1,00	1,00	9	225	0	0	9
G, R, U, X	1,00	0,78	0,88	7	225	0	2	9
H, P	0,75	1,00	0,86	9	222	3	0	9
I, N	1,00	0,67	0,80	6	225	0	3	9
A	0,53	1,00	0,69	9	217	8	0	9
B	0,69	1,00	0,82	9	221	4	0	9
C	0,57	0,89	0,70	8	219	6	1	9
E	0,86	0,67	0,75	6	224	1	3	9
F	0,90	1,00	0,95	9	224	1	0	9
J	0,82	1,00	0,90	9	223	2	0	9
M	1,00	0,22	0,36	2	225	0	7	9
O	0,62	0,56	0,59	5	222	3	4	9
Q	0,75	0,67	0,71	6	223	2	3	9
S	0,60	0,67	0,63	6	221	4	3	9
T	0,78	0,78	0,78	7	223	2	2	9
W	0,71	0,56	0,63	5	223	2	4	9
Z	0,89	0,89	0,89	8	224	1	4	9

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 475-490

Tabel 9. hasil setiap *DeiT-Tiny* kelas beserta nilai TP, TN, FP, dan FN

Kelas	Precision	Recall	F1-score	TP	TN	FP	FN	Sampel
A, B, D, E, G, H, J, K, L, M, N, P, Q, R, S, T, U, V, W, X, Y	1,00	1,00	1,00	9	225	0	0	9
C, F	0,82	1,00	0,90	9	223	2	0	9
I, Z	1,00	0,78	0,88	7	225	0	2	9
O	1,00	0,67	0,80	6	225	0	0	9

Perhitungan TP, TN, FP, dan FN dilakukan menggunakan pendekatan one-vs-rest. Ar*Tiny*a, setiap huruf secara bergantian dianggap sebagai kelas positif, sedangkan 25 huruf lainnya dianggap sebagai kelas negatif. Oleh karena itu, nilai tersebut harus dibaca relatif terhadap kelas target dan tidak dapat langsung dijumlahkan sebagai *confusion matrix* biner tunggal.



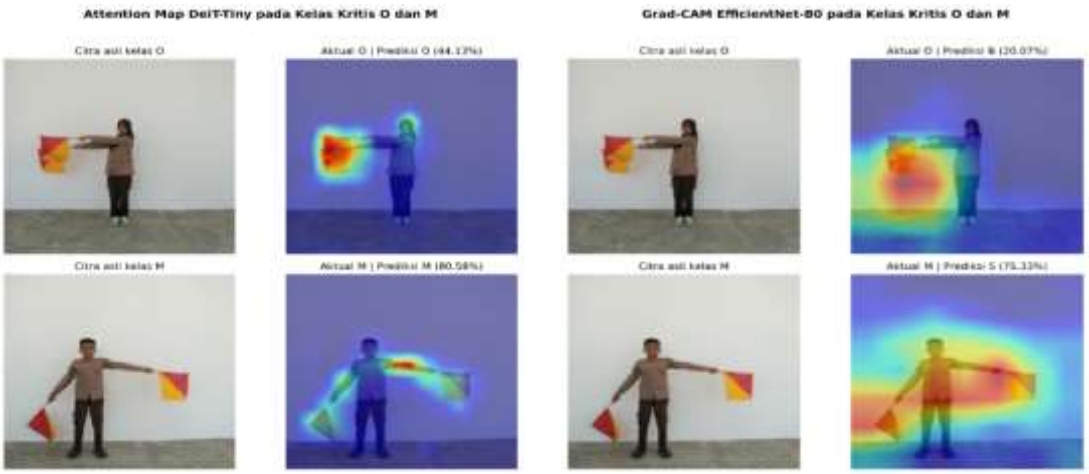
Gambar 3. Perbandingan Akurasi per Kelas

Nilai yang ditampilkan pada Gambar 3 dihitung dari jumlah prediksi benar dibagi jumlah citra aktual pada setiap kelas. Dengan demikian, ukuran tersebut secara matematis setara dengan recall per kelas. *EfficientNetB0* memperoleh nilai 100% pada sembilan kelas, sedangkan *DeiT-Tiny* mencapainya pada 23 kelas. Perbedaan paling besar terlihat pada kelas M. *EfficientNetB0* hanya mencapai 22,22%, sedangkan *DeiT-Tiny* mencapai 100%. Pada kelas O, *EfficientNetB0* memperoleh 55,56% dan *DeiT-Tiny* memperoleh 66,67%. Hasil tersebut menunjukkan bahwa kelas O tetap menjadi tantangan bagi kedua model, sedangkan kelemahan *EfficientNetB0* tersebar pada lebih banyak kelas.

Analisis Interpretabilitas Model Grad-CAM dan Attention Map

Untuk memahami bagian citra mana yang menjadi dasar keputusan tiap arsitektur, dilakukan analisis interpretabilitas menggunakan Grad-CAM pada *EfficientNetB0* (CNN) dan peta atensi (attention map) dari token [CLS] pada *DeiT-Tiny* (ViT). Analisis difokuskan pada dua kelas kritis, yaitu O dan M, karena kedua kelas tersebut merupakan sumber kesalahan utama pada EfficientNet-B0. Sampel yang ditampilkan dipilih pada kondisi *DeiT-Tiny* memprediksi benar sementara *EfficientNetB0* memprediksi salah, sehingga perbedaan fokus fitur kedua model dapat dibandingkan secara langsung.

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 475-490



Gambar 4. Visualisasi Grad-CAM dan Attention Map

Berdasarkan Gambar X dan Y, terlihat perbedaan pola fokus yang jelas antara kedua arsitektur. Pada kelas O, peta Grad-CAM *EfficientNetB0* cenderung teraktivasi pada area yang tersebar termasuk sebagian latar belakang dan badan subjek sehingga model gagal menangkap konfigurasi kedua lengan secara utuh dan salah mengklasifikasikan citra sebagai huruf B dengan keyakinan rendah (20,07%). Sebaliknya, peta atensi *DeiT-Tiny* terkonsentrasi pada posisi kedua bendera dan sudut lengan, yaitu fitur yang benar-benar menentukan makna huruf, sehingga model memprediksi O dengan benar.

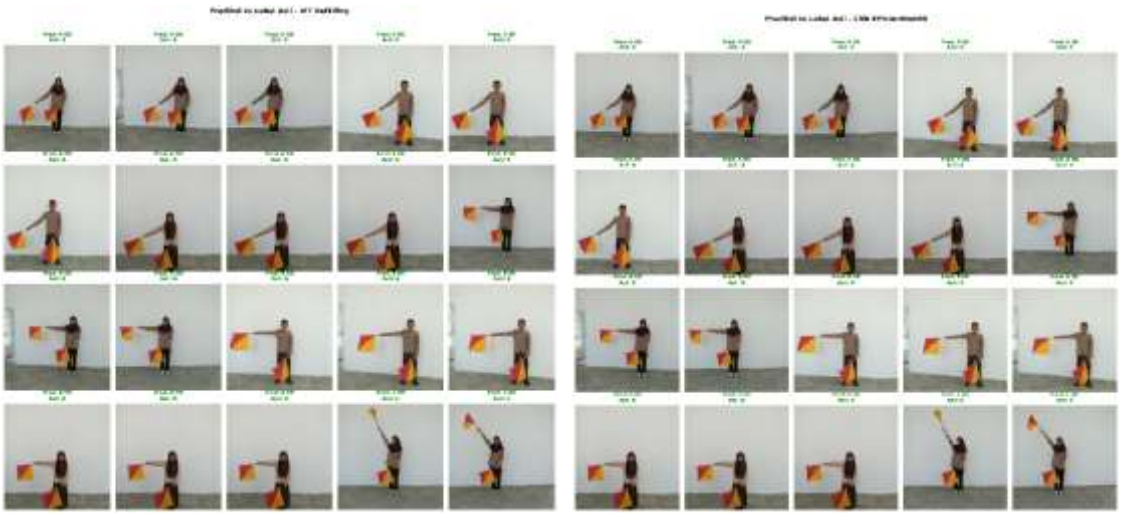
Tabel 10. Ringkasan Analisis pada Kelas Kritis (O dan M)

Kelas	Prediksi *EfficientNetB0* (Grad-CAM)	Keyakinan	Prediksi *DeiT-Tiny* (*Attention*)	Keyakinan
O	B *(salah)*	20,07%	O *(benar)*	44,13%
M	S *(salah)*	75,33%	M *(benar)*	80,58%

Pola serupa tampak pada kelas M. *EfficientNetB0* memusatkan aktivasi pada salah satu sisi lengan saja dan salah mengklasifikasikan citra sebagai huruf S meskipun dengan keyakinan tinggi (75,33%) indikasi bahwa model yakin pada fitur yang keliru. *DeiT-Tiny*, melalui mekanisme self-attention, mampu mengaitkan kedua lengan secara global dan memprediksi M dengan benar (keyakinan 80,58%).

Temuan interpretabilitas ini memberikan penjelasan komputasional atas keunggulan *DeiT-Tiny*: mekanisme self-attention memungkinkan model menangkap relasi spasial global antara posisi kedua lengan dan bendera, sedangkan receptive field lokal pada *EfficientNetB0* cenderung terpaku pada fitur parsial dan mudah terganggu oleh variasi latar belakang. Hal ini konsisten dengan analisis inductive bias pada Pendahuluan dan menjelaskan mengapa kesalahan *EfficientNetB0* tersebar pada lebih banyak kelas dengan pose yang mirip secara lokal.

Journal : Journal of Computer Science and Information Technology (JCSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : 3031-8467
Volume / Issue / Year/ Page : Vol. 3 No. 4 (2026): September/ Hal 475-490



Gambar 5. Visualisasi Hasil Prediksi

Visualisasi menampilkan citra pengujian beserta label aktual dan hasil prediksi model. Warna hijau menunjukkan prediksi benar, sedangkan warna merah menunjukkan prediksi yang tidak sesuai. Gambar tersebut hanya menggunakan 20 citra pertama berdasarkan urutan *Dataset* sehingga berfungsi sebagai ilustrasi, bukan sebagai dasar perhitungan keseluruhan metrik. Evaluasi kuantitatif tetap menggunakan seluruh 234 citra pengujian.

Secara keseluruhan, *DeiT-Tiny* memberikan hasil yang lebih baik dibandingkan *EfficientNetB0*. Keunggulan tersebut terlihat sejak proses validasi, yaitu melalui penurunan validation loss yang lebih cepat, akurasi validasi yang lebih tinggi, dan jarak pelatihan-validasi yang lebih kecil. Pola yang sama berlanjut pada data pengujian eksternal, ketika jumlah kesalahan menurun dari 42 menjadi tujuh citra.

Hasil tersebut menunjukkan bahwa pada Dataset penelitian ini, *DeiT-Tiny* lebih efektif dalam mempelajari hubungan spasial antarbagi¬an citra yang membentuk pose *semaphore*. Namun, hasil eksperimen tidak cukup untuk menyimpulkan bahwa mekanisme self-attention merupakan satu-satunya penyebab keunggulan tersebut. Perbedaan jumlah parameter, dinamika optimasi, bobot pralatih, dan karakteristik Dataset juga dapat memengaruhi hasil.

EfficienNetB0 tetap mampu memperoleh accuracy 82,05%, tetapi menunjukkan indikasi overfitting yang lebih kuat dan kesalahan yang tersebar pada lebih banyak kelas. Sebaliknya, *DeiT-Tiny* memperoleh accuracy 97,01%, F1-score 96,95%, serta prediksi sempurna pada 23 kelas. Berdasarkan seluruh bukti validasi dan pengujian, *DeiT-Tiny* menjadi model terbaik untuk klasifikasi huruf *semaphore* pada protokol eksperimen ini.

KESIMPULAN

Penelitian ini membandingkan kinerja Convolutional Neural Network (EfficientNet-B0) dan Vision Transformer (*DeiT-Tiny*) dalam klasifikasi citra 26 huruf *semaphore* Pramuka menggunakan konfigurasi pelatihan yang setara dan evaluasi pada data uji eksternal. Berdasarkan hasil eksperimen, *DeiT-Tiny* memberikan kinerja yang lebih unggul dibandingkan *EfficientNetB0* pada seluruh metrik, dengan accuracy 97,01%, precision 97,64%, recall 97,01%, dan F1-score 96,95%, serta mampu mengklasifikasikan 23 dari 26 kelas secara sempurna dengan hanya tujuh kesalahan. Sebaliknya, *EfficientNetB0* memperoleh accuracy 82,05% dan F1-score 81,53% dengan 42 kesalahan yang tersebar pada lebih

banyak kelas serta menunjukkan indikasi overfitting yang lebih kuat. Keunggulan *DeiT-Tiny* telah terlihat sejak tahap validasi melalui penurunan validation loss yang lebih cepat dan kestabilan akurasi yang lebih tinggi. Analisis interpretabilitas (Grad-CAM dan attention map) memperkuat temuan tersebut, yaitu bahwa *DeiT-Tiny* lebih fokus pada konfigurasi posisi kedua lengan dan bendera sebagai fitur penentu kelas, sedangkan *EfficientNetB0* cenderung terpaku pada fitur lokal parsial dan variasi latar belakang. Dengan demikian, pada dataset yang digunakan, *DeiT-Tiny* merupakan model terbaik untuk klasifikasi huruf *semaphore* Pramuka berbasis citra dan berpotensi menjadi alternatif yang lebih baik dalam pengembangan sistem pengenalan *semaphore*.

Meskipun demikian, penelitian ini memiliki beberapa keterbatasan. Dataset dikumpulkan pada kondisi yang relatif terkendali—dengan latar belakang seragam, pencahayaan stabil, dan jumlah subjek terbatas (13 subjek)—sehingga ketahanan model terhadap kondisi pencahayaan ekstrem, latar belakang kompleks, serta variasi sudut dan jarak pengambilan gambar yang lebih beragam belum sepenuhnya teruji dan berpotensi menurunkan kinerja pada penerapan di dunia nyata. Untuk penelitian selanjutnya (*future works*), disarankan untuk memperluas dataset dengan variasi kondisi lingkungan yang lebih realistis, serta melakukan optimasi dan implementasi model *DeiT-Tiny* pada perangkat *edge/embedded system* (misalnya melalui kuantisasi atau *knowledge distillation*) agar dapat digunakan untuk pengenalan huruf *semaphore* secara *real-time* di lapangan.

Pernyataan Mengenai Penggunaan Kecerdasan Buatan (AI)

Dalam proses penyusunan naskah artikel ini, penulis menggunakan teknologi kecerdasan buatan (Artificial Intelligence) secara terbatas sebagai alat bantu untuk memperbaiki struktur bahasa, menyempurnakan susunan kalimat, dan meningkatkan kejelasan penyampaian informasi. Penggunaan AI tidak menggantikan proses penelitian, pengumpulan dan pengolahan dataset, pelatihan model EfficientNetB0 dan DeiT-Tiny, maupun analisis hasil eksperimen. Seluruh bagian naskah yang memperoleh bantuan AI telah diperiksa, diverifikasi, dan disesuaikan kembali oleh penulis dengan hasil penelitian yang sebenarnya. Penulis bertanggung jawab sepenuhnya terhadap keakuratan, keaslian, dan integritas seluruh isi artikel.

Konflik Kepentingan

Penulis menyatakan bahwa tidak terdapat konflik kepentingan, baik dalam bentuk finansial, institusional, maupun personal, yang dapat memengaruhi pelaksanaan penelitian, interpretasi hasil, ataupun proses publikasi artikel ini.

Kontribusi Penulis

Muh. Akbar Haeruddin bertanggung jawab atas pelaksanaan utama penelitian, meliputi perancangan penelitian, pengumpulan dan pengolahan dataset citra huruf Semaphore, implementasi model EfficientNetB0 dan DeiT-Tiny, pelatihan dan pengujian model, analisis hasil evaluasi, serta penyusunan naskah artikel. Desi Anggreani dan Muhyiddin A.M. Hayat berperan dalam memberikan arahan dan bimbingan selama proses penelitian, meninjau hasil eksperimen, serta memberikan masukan untuk penyempurnaan isi dan kualitas naskah. Seluruh penulis telah membaca, meninjau, dan menyetujui versi akhir artikel untuk diterbitkan.

Ucapan Terima Kasih

Penulis menyampaikan terima kasih kepada Universitas Muhammadiyah

Makassar, khususnya Fakultas Teknik dan Program Studi Informatika, atas dukungan akademik dan fasilitas yang diberikan selama pelaksanaan penelitian. Ucapan terima kasih juga disampaikan kepada dosen pembimbing atas arahan, bimbingan, dan masukan yang diberikan selama proses penelitian dan penyusunan artikel. Penulis turut menyampaikan apresiasi kepada seluruh pihak yang telah membantu dalam proses pengumpulan dataset citra huruf Semaphore serta memberikan dukungan hingga penelitian ini dapat diselesaikan dengan baik.

DAFTAR PUSTAKA

[1] K. Raman, W. W. Lun, and H. Hashim, "Virtual reality for verbal communication development in English as second language learning: Advantages and optimisation strategies," *Comput. Educ. X Real.*, vol. 8, no. February, p. 100145, Jun. 2026.

[2] Y. Huan and W. Yan, "Semaphore Recognition Using Deep Learning," *Electron.*, vol. 14, no. 2, 2025.

[3] R. Jalayer, M. Jalayer, C. Orsenigo, and M. Tomizuka, "A review on deep learning for vision-based hand detection, hand segmentation and hand gesture recognition in human–robot interaction," *Robot. Comput. Integr. Manuf.*, vol. 97, no. July 2025, p. 103110, 2026.

[4] P. Niu, "Convolutional neural network for gesture recognition human-computer interaction system design," *PLoS One*, vol. 20, no. 2 February, pp. 1–22, 2025.

[5] Purwono, A. Ma'arif, W. Rahmaniar, H. I. K. Fathurrahman, A. Z. K. Frisky, and Q. M. U. Haq, "Understanding of Convolutional Neural Network (CNN): A Review," *Int. J. Robot. Control Syst.*, vol. 2, no. 4, pp. 739–748, 2022.

[6] G. Rangel, J. C. Cuevas-Tello, J. Nunez-Varela, C. Puente, and A. G. Silva-Trujillo, "A Survey on Convolutional Neural Networks and Their Performance Limitations in Image Recognition Tasks," *Journal of Sensors*, vol. 2024. Hindawi Limited, 2024.

[7] L. S. A. Putra, F. T. P. Wigyarinto, E. Kusumawardhani, and V. A. Gunawan, "Semaphore letter code recognition system using wavelet method and back propagation neural network," *Int. J. Adv. Technol. Eng. Explor.*, vol. 11, no. 112, pp. 316–331, 2024.

[8] J. Maurício, I. Domingues, and J. Bernardino, "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review," *Appl. Sci.*, vol. 13, no. 9, 2023.

[9] T. Zhang, W. Xu, B. Luo, and G. Wang, "Depth-Wise Convolutions in Vision Transformers for efficient training on small datasets," *Neurocomputing*, vol. 617, no. March 2024, p. 128998, 2025.

[10] C. K. Tan, K. M. Lim, R. K. Y. Chang, C. P. Lee, and A. Alqahtani, "HGR-ViT: Hand Gesture Recognition with Vision Transformer," *Sensors*, vol. 23, no. 12, pp. 1–20, 2023.