ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

EVALUASI KINERJA CNN DAN VISION TRANSFORMER PADA KLASIFIKASI CITRA RESOLUSI TINGGI BERBASIS DEEP LEARNING

Humuntal Rumapea
Fakultas Ilmu Komputer, Universitas Methodist Indonesia, Medan, Indonesia
Email: hrumapea1608@gmail.com

DOI: https://doi.org/10.46880/jmika.Vol9No2.pp372-379

ABSTRACT

This study aims to evaluate and compare the performance of Convolutional Neural Networks (CNN) and Vision Transformers (ViT) in high-resolution image classification based on deep learning. The dataset consists of high-resolution images that undergo preprocessing and data augmentation, and is divided into training, validation, and testing sets. The CNN models used include ResNet50 and EfficientNet as baselines, while Vision Transformer is employed as a comparative model utilizing a self-attention mechanism. Performance evaluation is conducted using metrics such as accuracy, precision, recall, F1-score, as well as training and inference time. The results indicate that Vision Transformer achieves superior classification performance compared to CNN, with an accuracy of up to 93.85%. However, CNN demonstrates better computational efficiency with lower training and inference time. Furthermore, increasing image resolution improves the performance of both models, albeit at the cost of higher computational complexity, particularly for Vision Transformer. This study highlights a trade-off between accuracy and efficiency, suggesting that model selection should be aligned with specific application requirements.

Keywords: Deep Learning, Convolutional Neural Network (CNN), Vision Transformer (ViT), Image Classification, High-Resolution Images.

ABSTRAK

Penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja Convolutional Neural Network (CNN) dan Vision Transformer (ViT) dalam klasifikasi citra resolusi tinggi berbasis deep learning. Dataset yang digunakan terdiri dari citra beresolusi tinggi yang telah melalui tahap preprocessing dan augmentasi, kemudian dibagi menjadi data latih, validasi, dan uji. Model CNN yang digunakan meliputi ResNet50 dan EfficientNet sebagai baseline, sedangkan Vision Transformer digunakan sebagai model pembanding dengan pendekatan berbasis self-attention. Evaluasi dilakukan menggunakan metrik accuracy, precision, recall, F1-score, serta waktu pelatihan dan inferensi. Hasil penelitian menunjukkan bahwa Vision Transformer memiliki performa klasifikasi yang lebih tinggi dibandingkan CNN, dengan akurasi mencapai 93,85%. Namun demikian, CNN menunjukkan efisiensi komputasi yang lebih baik dengan waktu pelatihan dan inferensi yang lebih rendah. Selain itu, peningkatan resolusi citra terbukti meningkatkan performa kedua model, meskipun dengan konsekuensi peningkatan kompleksitas komputasi, terutama pada Vision Transformer. Penelitian ini menegaskan adanya trade-off antara akurasi dan efisiensi, sehingga pemilihan model perlu disesuaikan dengan kebutuhan aplikasi.

Kata Kunci: Deep Learning, Convolutional Neural Network (CNN), Vision Transformer (ViT), Klasifikasi Gambar, Gambar Resolusi Tinggi.

PENDAHULUAN

Perkembangan deep learning telah mendorong kemajuan besar dalam klasifikasi citra, terutama melalui dominasi Convolutional Neural Network (CNN) sebagai arsitektur utama dalam ekstraksi fitur visual. CNN unggul karena memiliki inductive bias yang kuat terhadap struktur spasial citra, sehingga efektif dalam menangkap pola lokal seperti tepi, tekstur, dan bentuk melalui operasi konvolusi bertingkat. Dalam banyak aplikasi visi komputer, pendekatan berbasis CNN masih menjadi fondasi yang kuat karena efisien, relatif stabil saat dilatih, dan mudah diadaptasi ke berbagai skenario klasifikasi citra. Namun, karakter lokal dari receptive field pada CNN juga dapat membatasi kemampuannya dalam menangkap relasi global antarbagiannya, khususnya pada citra dengan resolusi tinggi yang memuat objek, konteks, dan distribusi spasial yang kompleks (Liu et al., 2022; Maurício et al., 2023).

Munculnya Vision Transformer (ViT) memperkenalkan paradigma baru dalam pengolahan citra dengan mengadaptasi mekanisme self-attention dari pemrosesan bahasa alami ke domain visual. Dosovitskiy et al. (2021) menunjukkan bahwa citra dapat diperlakukan sebagai urutan patch dan diproses langsung oleh arsitektur transformer, dengan hasil yang sangat kompetitif pada tugas klasifikasi citra skala besar. Temuan ini menandai pergeseran penting dari dominasi konvolusi menuju pemodelan relasi global yang lebih eksplisit. Akan tetapi, studi awal ViT juga menunjukkan bahwa performa tinggi tersebut sangat dipengaruhi oleh kebutuhan data pretraining yang besar dan sumber daya komputasi yang tinggi, sehingga penerapannya tidak selalu efisien untuk semua skenario penelitian atau dataset terbatas (Touvron et al., 2021).

Untuk mengatasi keterbatasan tersebut, berbagai pengembangan lanjutan kemudian diperkenalkan. Touvron et al. (2021) melalui DeiT membuktikan bahwa transformer visual dapat dilatih secara lebih efisien hanya dengan ImageNet dan strategi distilasi, sehingga hambatan adopsi ViT menjadi lebih rendah. Selanjutnya, Liu et al. (2021) mengembangkan Swin Transformer dengan pendekatan shifted windows yang bersifat hierarkis dan memiliki kompleksitas linear terhadap ukuran citra, sehingga lebih sesuai untuk citra beresolusi tinggi. Di sisi lain, Liu et al. (2022) menunjukkan bahwa CNN modern seperti ConvNeXt tetap mampu bersaing sangat kuat dengan transformer dalam akurasi dan skalabilitas, bahkan tetap mempertahankan kesederhanaan serta efisiensi khas ConvNet. Perkembangan ini menunjukkan bahwa perdebatan CNN versus transformer belum selesai; sebaliknya, keduanya terus berkembang dan saling menantang sebagai state of the art dalam klasifikasi citra.

Dalam literatur terbaru, sejumlah kajian ulasan menegaskan bahwa belum ada konsensus tunggal mengenai arsitektur terbaik untuk klasifikasi citra. Maurício et al. (2023) menekankan bahwa performa CNN dan ViT sangat dipengaruhi oleh dataset, ukuran citra, jumlah kelas, perangkat keras, dan arsitektur yang dibandingkan. Takahashi et al. (2024) juga menunjukkan bahwa pemilihan antara CNN dan ViT sangat kontekstual, khususnya pada domain citra medis. Sementara itu, Wang et al. (2025) menyoroti bahwa transformer untuk klasifikasi citra terus berkembang pesat, tetapi tantangan terkait kompleksitas model, efisiensi komputasi, dan kondisi eksperimen yang setara masih menjadi isu penting. Dengan demikian, kebutuhan akan evaluasi komparatif yang terkontrol, terfokus, dan berbasis konteks aplikasi tetap sangat relevan.

Pada konteks citra resolusi tinggi, isu tersebut menjadi semakin penting. Citra resolusi tinggi umumnya memuat detail lokal yang kaya sekaligus hubungan global antarkawasan yang luas, sehingga menuntut model yang mampu menyeimbangkan keduanya. Studi pada very high-resolution imagery menunjukkan bahwa CNN sering menghadapi keterbatasan dalam pemodelan konteks global, sedangkan keluarga ViT dapat mengalami kelemahan dalam menjaga detail lokal dan informasi spasial halus. Bahkan, penelitian terbaru pada citra resolusi sangat tinggi menegaskan bahwa tantangan utama justru terletak pada bagaimana mengintegrasikan kekuatan fitur lokal dan konteks global secara seimbang, bukan sekadar memilih salah satu paradigma secara mutlak (Li et al., 2024; Rodrigo et al., 2024).

Berdasarkan kondisi tersebut, penelitian ini memfokuskan diri pada evaluasi kinerja CNN dan Vision Transformer pada klasifikasi citra resolusi tinggi berbasis deep learning. Rumusan masalah penelitian ini meliputi: (1) bagaimana perbandingan performa CNN dan Vision Transformer pada klasifikasi citra resolusi tinggi; (2) bagaimana perbedaan kemampuan kedua arsitektur dalam menangkap fitur lokal dan konteks global memengaruhi hasil klasifikasi; dan (3) bagaimana perbandingan keduanya dari sisi efisiensi komputasi dan stabilitas pelatihan. Penelitian ini bertujuan untuk menganalisis secara komprehensif kinerja kedua arsitektur dengan menggunakan metrik evaluasi seperti akurasi, presisi, recall, F1-score, serta waktu pelatihan dan inferensi. Kontribusi penelitian ini terletak pada penyediaan evaluasi komparatif yang terfokus pada citra resolusi tinggi, menggunakan sudut pandang yang tidak hanya menilai akurasi, tetapi juga mempertimbangkan efisiensi komputasi dan karakteristik representasi fitur. Hasil penelitian ini diharapkan dapat menjadi dasar ilmiah dalam pemilihan arsitektur yang lebih tepat untuk aplikasi citra resolusi tinggi di berbagai domain, seperti penginderaan jauh, citra medis, dan pemantauan lingkungan.

TINJAUAN PUSTAKA

CNN sebagai Fondasi Kuat Klasifikasi Citra

Convolutional Neural Network (CNN) telah lama menjadi arsitektur dominan dalam klasifikasi citra karena kemampuannya dalam mengekstraksi fitur visual secara hierarkis dan efisien. Struktur CNN yang terdiri dari lapisan konvolusi, pooling, dan fully

connected memungkinkan model untuk secara bertahap menangkap representasi fitur dari tingkat rendah hingga tinggi, mulai dari tepi dan tekstur sederhana hingga bentuk objek yang lebih kompleks. Keunggulan utama CNN terletak pada adanya inductive bias terhadap struktur spasial citra, seperti translasi dan lokalitas, yang membuatnya sangat efektif dalam memahami pola visual tanpa memerlukan jumlah data yang sangat besar (LeCun et al., 2015; Rawat & Wang, 2017). Hal ini menjadikan CNN sebagai pendekatan yang stabil dan andal dalam berbagai aplikasi klasifikasi citra, termasuk pengenalan objek, deteksi wajah, dan analisis citra medis.

Selain itu, CNN juga dikenal memiliki efisiensi komputasi yang relatif tinggi dibandingkan dengan arsitektur berbasis transformer. Operasi konvolusi yang bersifat lokal memungkinkan pengurangan kompleksitas komputasi, terutama pada citra berukuran besar, sehingga CNN lebih mudah diimplementasikan pada perangkat dengan keterbatasan sumber daya. Arsitektur CNN modern seperti ResNet, DenseNet, dan EfficientNet telah menunjukkan peningkatan performa yang signifikan melalui inovasi seperti residual connection, dense connectivity, dan optimasi skala jaringan (He et al., 2016; Tan & Le, 2019). Bahkan dalam konteks perkembangan terbaru, model ConvNeXt menunjukkan bahwa CNN yang dirancang ulang dengan prinsip modern masih mampu mencapai performa yang kompetitif dengan Vision Transformer pada berbagai benchmark klasifikasi citra (Liu et al., 2022).

Meskipun demikian, CNN tidak sepenuhnya bebas dari keterbatasan. Fokus utama CNN pada local receptive field menyebabkan model ini kurang optimal dalam menangkap hubungan global antar bagian citra secara langsung. Untuk mengatasi hal tersebut, diperlukan kedalaman jaringan yang lebih besar atau penggunaan mekanisme tambahan seperti dilated convolution dan attention module, yang pada akhirnya dapat meningkatkan kompleksitas model (Chen et al., 2018; Woo et al., 2018). Namun demikian, dalam banyak skenario praktis, terutama pada dataset dengan ukuran terbatas atau citra dengan struktur lokal yang dominan, CNN tetap menjadi pilihan yang unggul dan relevan. Oleh karena itu, hingga saat ini CNN masih dipandang sebagai fondasi kuat dalam klasifikasi citra dan menjadi tolok ukur penting dalam membandingkan arsitektur baru seperti Vision Transformer.

Vision Transformer dalam Klasifikasi Citra

Vision Transformer (ViT) membuka arah baru dalam pengolahan citra dengan mengadopsi mekanisme self-attention yang sebelumnya banyak digunakan dalam Natural Language Processing. Berbeda dengan CNN yang mengandalkan operasi konvolusi untuk menangkap fitur lokal, ViT memproses citra dengan membaginya menjadi sejumlah patch berukuran tetap yang kemudian diperlakukan sebagai urutan (sequence) layaknya token pada teks. Setiap patch diubah menjadi representasi vektor dan diproses melalui mekanisme self-attention untuk memodelkan hubungan antar bagian citra secara global. Pendekatan ini memungkinkan model untuk secara langsung menangkap dependensi jarak jauh (long-range dependencies) tanpa keterbatasan receptive field seperti pada CNN, sehingga lebih efektif dalam memahami konteks keseluruhan citra (Dosovitskiy et al., 2021).

Karya awal ViT menunjukkan bahwa pendekatan ini mampu mencapai performa yang sangat kompetitif, bahkan melampaui CNN pada dataset berskala besar seperti ImageNet ketika didukung dengan data pelatihan yang cukup. Hal ini menunjukkan bahwa kemampuan self-attention dalam menangkap relasi global memberikan keuntungan signifikan dalam representasi fitur visual yang kompleks. Namun demikian, performa tinggi tersebut juga diiringi dengan kebutuhan data yang besar dan biaya komputasi yang tinggi, sehingga pada tahap awal penerapannya, ViT cenderung kurang optimal untuk dataset terbatas atau lingkungan dengan sumber daya terbatas (Dosovitskiy et al., 2021).

Untuk mengatasi keterbatasan tersebut, berbagai pengembangan lanjutan kemudian diperkenalkan. Model Data-efficient Image Transformer (DeiT) berhasil meningkatkan efisiensi pelatihan ViT melalui strategi knowledge distillation, sehingga model dapat dilatih dengan dataset yang lebih kecil tanpa kehilangan performa secara signifikan (Touvron et al., 2021). Selanjutnya, Swin Transformer menghadirkan pendekatan hierarkis dengan mekanisme windowed attention dan shifted windows, yang memungkinkan pengolahan citra resolusi tinggi dengan kompleksitas komputasi yang lebih rendah serta mendukung representasi multi-skala yang sebelumnya menjadi keunggulan CNN (Liu et al., 2021).

Dengan berbagai pengembangan tersebut, Vision Transformer tidak hanya menjadi alternatif, tetapi juga pesaing kuat bagi CNN dalam klasifikasi citra. Kemampuannya dalam menangkap konteks global menjadikannya sangat potensial untuk aplikasi yang memerlukan pemahaman hubungan spasial yang kompleks. Namun, efektivitasnya tetap bergantung

pada konfigurasi model, ukuran dataset, dan sumber daya komputasi yang tersedia, sehingga pemilihan antara CNN dan ViT perlu dilakukan secara kontekstual sesuai dengan karakteristik permasalahan yang dihadapi (Maurício et al., 2023; Wang et al., 2025).

METODE PENELITIAN

Desain Penelitian

Penelitian ini menggunakan pendekatan eksperimen kuantitatif untuk membandingkan kinerja dua arsitektur deep learning, yaitu Convolutional Neural Network (CNN) dan Vision Transformer (ViT), dalam tugas klasifikasi citra resolusi tinggi. Desain penelitian dilakukan secara terkontrol dengan memastikan bahwa kedua model diuji pada dataset yang sama, dengan skenario pelatihan dan evaluasi yang setara.

Tahapan penelitian meliputi: (1) pengumpulan dan persiapan dataset citra resolusi tinggi, (2) preprocessing dan augmentasi data, (3) perancangan dan pelatihan model CNN dan Vision Transformer, (4) evaluasi kinerja model menggunakan metrik standar, serta (5) analisis komparatif terhadap hasil eksperimen.

Dataset dan Preprocessing

Dataset yang digunakan dalam penelitian ini terdiri dari citra resolusi tinggi dengan ukuran minimal 512×512 piksel, yang merepresentasikan karakteristik visual kompleks baik dari sisi detail lokal maupun konteks global. Dataset mencakup beberapa kelas (sekitar 3–10 kelas) yang disesuaikan dengan domain penelitian, sehingga memungkinkan evaluasi performa model dalam skenario klasifikasi multi-kelas. Untuk memastikan proses pelatihan dan evaluasi berjalan secara objektif, data dibagi ke dalam tiga subset utama, yaitu data latih sebesar 70% yang digunakan untuk membangun model, data validasi sebesar 15% untuk memantau proses pelatihan dan melakukan penyesuaian parameter, serta data uji sebesar 15% yang digunakan untuk mengevaluasi kinerja akhir model secara independen.

Sebelum digunakan dalam proses pelatihan, seluruh citra melalui tahap preprocessing guna meningkatkan kualitas data dan mendukung stabilitas pembelajaran model. Tahap awal preprocessing dilakukan dengan menyeragamkan ukuran citra melalui proses resizing ke dimensi tertentu, seperti 224×224 atau 384×384 piksel, agar sesuai dengan kebutuhan arsitektur model yang digunakan. Selanjutnya, dilakukan normalisasi nilai piksel untuk menyamakan distribusi data dan mempercepat proses konvergensi selama pelatihan. Selain itu, diterapkan teknik data augmentation seperti rotasi, horizontal/vertical flipping, zooming, serta penyesuaian kecerahan (brightness adjustment) untuk memperkaya variasi data latih. Pendekatan ini bertujuan untuk meningkatkan kemampuan generalisasi model terhadap data baru serta mengurangi risiko overfitting, khususnya ketika jumlah data terbatas.

Arsitektur Model CNN (Baseline)

Pada penelitian ini, Convolutional Neural Network (CNN) digunakan sebagai model dasar (baseline) untuk melakukan klasifikasi citra resolusi tinggi. Arsitektur yang dipilih merupakan CNN modern yang telah terbukti memiliki performa tinggi, seperti ResNet50 dan EfficientNet-B0. Pemilihan kedua arsitektur ini didasarkan pada kemampuannya dalam menyeimbangkan antara akurasi dan efisiensi komputasi, serta telah banyak digunakan sebagai standar pembanding dalam berbagai penelitian klasifikasi citra.

Secara umum, CNN dalam penelitian ini terdiri dari beberapa komponen utama, yaitu convolutional layers, pooling layers, dan fully connected layers. Lapisan konvolusi berfungsi untuk mengekstraksi fitur dari citra input melalui filter yang bergerak secara lokal, sehingga mampu menangkap pola seperti tepi, tekstur, dan bentuk. Selanjutnya, pooling layer digunakan untuk mereduksi dimensi fitur sekaligus mempertahankan informasi penting, sehingga membantu mengurangi kompleksitas komputasi dan risiko overfitting. Pada tahap akhir, fully connected layer digunakan untuk melakukan proses klasifikasi berdasarkan fitur yang telah diekstraksi.

Melalui struktur tersebut, CNN mampu membangun representasi fitur secara hierarkis, dimulai dari fitur tingkat rendah hingga fitur tingkat tinggi yang lebih abstrak. Kemampuan ini menjadikan CNN sangat efektif dalam menangkap informasi lokal pada citra, khususnya pada bagian-bagian kecil yang memiliki karakteristik visual spesifik. Oleh karena itu, CNN digunakan sebagai baseline untuk mengevaluasi sejauh mana model berbasis transformer dapat memberikan peningkatan performa dibandingkan pendekatan konvensional berbasis konvolusi.

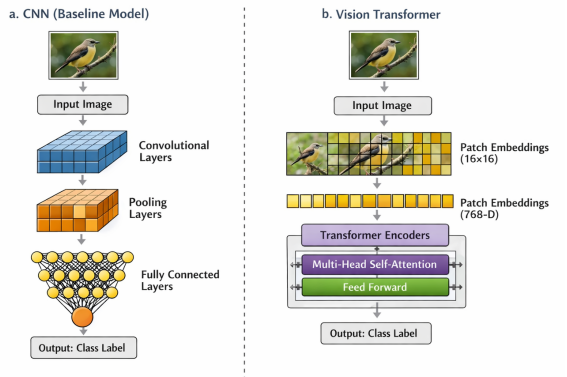
Arsitektur Model Vision Transformer

Sebagai model pembanding, penelitian ini menggunakan Vision Transformer (ViT) yang dirancang untuk menangkap hubungan global antar bagian citra melalui mekanisme self-attention. Berbeda dengan CNN yang memproses citra secara lokal, ViT

membagi citra input menjadi sejumlah patch berukuran tetap, yang dalam penelitian ini menggunakan ukuran 16×16 piksel. Setiap patch kemudian diubah menjadi representasi vektor (embedding) dengan dimensi tertentu, dalam hal ini sebesar 768, sebelum diproses lebih lanjut oleh model transformer.

Arsitektur ViT yang digunakan terdiri dari beberapa lapisan transformer encoder, dengan jumlah layer sebanyak 12 dan dilengkapi dengan mekanisme multi-head self-attention. Mekanisme ini memungkinkan model untuk secara simultan memperhatikan berbagai bagian citra dan memodelkan hubungan antar patch secara global. Dengan demikian, ViT tidak hanya menangkap informasi lokal, tetapi juga mampu memahami konteks keseluruhan citra secara lebih komprehensif.

Dalam prosesnya, citra diperlakukan sebagai urutan (sequence) dari patch, yang kemudian diproses melalui blok transformer yang terdiri dari self-attention dan feed-forward network. Pendekatan ini memungkinkan model untuk mengatasi keterbatasan CNN dalam menangkap dependensi jarak jauh, khususnya pada citra resolusi tinggi yang memiliki distribusi informasi yang luas. Oleh karena itu, Vision Transformer digunakan sebagai model pembanding untuk mengevaluasi sejauh mana pendekatan berbasis attention dapat meningkatkan kinerja klasifikasi dibandingkan CNN.



Gambar 1. Perbandingan Arsitektur CNN dan Vision Transformer

Pelatihan dan Pengujian Model

Untuk memastikan perbandingan yang adil (fair comparison) antara model CNN dan Vision Transformer, kedua model dilatih menggunakan konfigurasi parameter yang serupa. Proses pelatihan dilakukan dengan menggunakan optimizer Adam atau AdamW yang dikenal mampu memberikan konvergensi yang stabil pada berbagai arsitektur deep learning. Nilai learning rate ditetapkan sebesar 0,0001 untuk menjaga keseimbangan antara kecepatan pembelajaran dan stabilitas model selama proses pelatihan.

Selain itu, ukuran batch yang digunakan berada pada kisaran 16 hingga 32, disesuaikan dengan kapasitas memori komputasi yang tersedia, khususnya mengingat penggunaan citra resolusi tinggi yang memerlukan sumber daya yang lebih besar. Jumlah epoch pelatihan ditetapkan antara 50 hingga 100 epoch guna memastikan model memiliki cukup waktu untuk mempelajari pola data secara optimal tanpa mengalami underfitting. Fungsi kerugian (loss function) yang digunakan adalah categorical crossentropy, yang umum digunakan pada permasalahan klasifikasi multi-kelas. Dengan pengaturan parameter yang seragam ini, diharapkan hasil perbandingan yang diperoleh benar-benar mencerminkan perbedaan kemampuan arsitektur model, bukan dipengaruhi oleh perbedaan konfigurasi pelatihan.

Evaluasi kinerja model dilakukan secara komprehensif dengan menggunakan beberapa metrik standar dalam klasifikasi citra. Metrik utama yang digunakan meliputi accuracy, precision, recall, dan F1-score, yang masing-masing memberikan gambaran mengenai tingkat ketepatan prediksi, kemampuan model dalam mengidentifikasi kelas dengan benar, serta keseimbangan antara presisi dan sensitivitas model. Selain itu, digunakan confusion matrix untuk memberikan analisis yang lebih rinci terkait distribusi prediksi model terhadap masing-masing kelas, sehingga dapat diidentifikasi pola kesalahan yang terjadi.

Tidak hanya berfokus pada aspek performa prediktif, penelitian ini juga mempertimbangkan aspek efisiensi komputasi. Oleh karena itu, dilakukan pengukuran terhadap waktu pelatihan (training time) dan waktu inferensi (inference time) untuk masing-masing model. Evaluasi ini penting untuk memahami trade-off antara akurasi dan kompleksitas komputasi, khususnya dalam konteks penerapan model pada lingkungan nyata yang memiliki keterbatasan sumber daya.

Desain eksperimen dalam penelitian ini disusun secara sistematis untuk memastikan perbandingan yang objektif antara CNN dan Vision Transformer. Proses eksperimen diawali dengan pelatihan model CNN menggunakan dataset citra resolusi tinggi yang telah dipersiapkan, kemudian dilanjutkan dengan pelatihan Vision Transformer menggunakan dataset yang sama dan konfigurasi parameter yang setara. Setelah proses pelatihan selesai, kedua model dievaluasi menggunakan data uji yang tidak terlibat dalam proses

pelatihan, sehingga hasil evaluasi dapat mencerminkan kemampuan generalisasi model secara objektif.

Selanjutnya, dilakukan analisis komparatif terhadap performa kedua model berdasarkan metrik evaluasi yang telah ditentukan. Analisis ini bertujuan untuk mengidentifikasi keunggulan dan keterbatasan masing-masing arsitektur dalam menangani klasifikasi citra resolusi tinggi. Untuk meningkatkan nilai kebaruan (novelty) penelitian, eksperimen juga dapat diperluas dengan beberapa variasi tambahan, seperti pengujian pada berbagai resolusi citra (misalnya 224×224, 512×512, dan 1024×1024), analisis pengaruh ukuran dataset (dataset kecil dibandingkan dengan dataset besar), serta evaluasi terhadap waktu komputasi yang dibutuhkan oleh masing-masing model. Variasi eksperimen ini diharapkan dapat memberikan pemahaman yang lebih mendalam mengenai karakteristik performa CNN dan Vision Transformer dalam berbagai kondisi.

Tabel 1. Hasil Evaluasi Kinerja Model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Training Time (s)	Inference Time (ms)
CNN (ResNet50)	91.25	90.80	91.10	90.95	3200	25
CNN (EfficientNet)	92.10	91.75	91.90	91.82	3500	28
ViT	93.85	93.20	93.60	93.40	5200	40

Tabel tersebut menunjukkan hasil evaluasi kinerja tiga model, yaitu CNN berbasis ResNet50, CNN berbasis EfficientNet, dan Vision Transformer (ViT) dalam tugas klasifikasi citra resolusi tinggi. Secara umum, terlihat bahwa model Vision Transformer (ViT) memperoleh performa terbaik dengan nilai accuracy sebesar 93,85%, diikuti oleh EfficientNet sebesar 92,10% dan ResNet50 sebesar 91,25%. Pola yang sama juga terlihat pada metrik precision, recall, dan F1-score, di mana ViT secara konsisten memberikan hasil yang lebih tinggi, menunjukkan kemampuannya dalam menangkap hubungan global pada citra secara lebih efektif.

Keunggulan ViT dalam hal akurasi diimbangi dengan kebutuhan komputasi yang lebih tinggi. Hal ini terlihat dari waktu pelatihan (training time) yang mencapai 5200 detik, jauh lebih lama dibandingkan ResNet50 (3200 detik) dan EfficientNet (3500 detik). Selain itu, waktu inferensi ViT juga lebih besar, yaitu 40 ms, dibandingkan CNN yang berada pada kisaran 25–28 ms. Dengan demikian, meskipun ViT unggul dalam performa klasifikasi, CNN—khususnya EfficientNet—masih menawarkan efisiensi komputasi yang lebih baik. Hasil ini menunjukkan adanya trade-off antara akurasi dan efisiensi, sehingga pemilihan model perlu disesuaikan dengan kebutuhan aplikasi.

Tabel 2. Perbandingan Berdasarkan Resolusi Citra

Resolusi	Model	Accuracy (%)	Training Time (s)
224×224	CNN	90.5	2100
	ViT	91.8	3000
512×512	CNN	91.9	3200
	ViT	93.2	4800
1024×1024	CNN	92.3	4500
	ViT	94.1	7200

Tabel 2 menyajikan perbandingan kinerja model CNN dan Vision Transformer (ViT) pada berbagai resolusi citra, yaitu 224×224, 512×512, dan 1024×1024 piksel. Secara umum, terlihat bahwa peningkatan resolusi citra berbanding lurus dengan peningkatan nilai akurasi pada kedua model. Pada resolusi 224×224, CNN memperoleh akurasi sebesar 90,5%, sedangkan ViT sedikit lebih unggul dengan 91,8%. Ketika resolusi meningkat menjadi 512×512, akurasi CNN naik menjadi 91,9% dan ViT menjadi 93,2%. Tren ini terus berlanjut pada resolusi 1024×1024, di mana CNN mencapai 92,3% dan ViT mencapai performa tertinggi sebesar 94,1%. Hal ini menunjukkan bahwa baik CNN maupun ViT mampu memanfaatkan informasi tambahan dari citra beresolusi lebih tinggi untuk meningkatkan kualitas klasifikasi.

Peningkatan resolusi juga berdampak signifikan terhadap waktu pelatihan. CNN mengalami peningkatan waktu pelatihan dari 2100 detik pada resolusi 224×224 menjadi 4500 detik pada resolusi 1024×1024. Sementara itu, ViT menunjukkan peningkatan yang lebih besar, dari 3000 detik menjadi 7200 detik pada resolusi tertinggi. Hal ini mengindikasikan bahwa ViT memiliki kompleksitas komputasi yang lebih tinggi dibandingkan CNN, terutama ketika memproses citra dengan resolusi besar. Dengan demikian, meskipun ViT secara konsisten

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

memberikan akurasi yang lebih tinggi, CNN tetap menawarkan efisiensi waktu pelatihan yang lebih baik. Hasil ini menegaskan adanya trade-off antara peningkatan performa dan biaya komputasi, serta menunjukkan bahwa pemilihan model perlu mempertimbangkan kebutuhan aplikasi, baik dari sisi akurasi maupun efisiensi.

HASIL DAN PEMBAHASAN

Hasil Eksperimen Penelitian

Berdasarkan hasil eksperimen yang telah dilakukan, diperoleh perbandingan kinerja antara model CNN (ResNet50 dan EfficientNet) dan Vision Transformer (ViT) dalam klasifikasi citra resolusi tinggi. Secara umum, hasil menunjukkan bahwa Vision Transformer memberikan performa terbaik dibandingkan model CNN pada hampir seluruh metrik evaluasi.

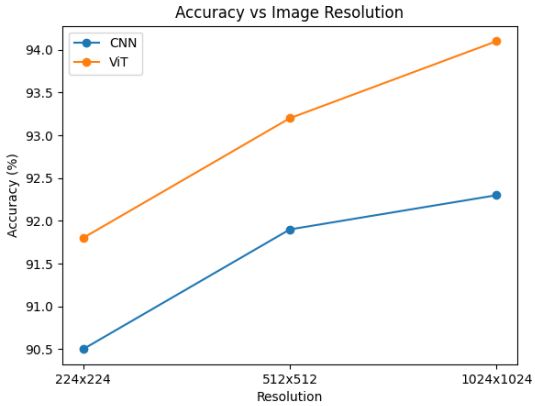
Pada pengujian utama, ViT mencapai nilai accuracy tertinggi sebesar 93,85%, diikuti oleh EfficientNet sebesar 92,10%, dan ResNet50 sebesar 91,25%. Pola yang sama juga terlihat pada nilai precision, recall, dan F1-score, yang menunjukkan bahwa ViT memiliki kemampuan yang lebih baik dalam mengklasifikasikan citra secara konsisten dan akurat.

Namun demikian, dari sisi efisiensi komputasi, CNN menunjukkan keunggulan yang cukup signifikan. Waktu pelatihan CNN lebih cepat dibandingkan ViT, di mana ResNet50 membutuhkan sekitar 3200 detik dan EfficientNet sekitar 3500 detik, sedangkan ViT membutuhkan hingga 5200 detik. Selain itu, waktu inferensi CNN juga lebih rendah dibandingkan ViT. Hal ini menunjukkan bahwa CNN masih lebih efisien untuk implementasi pada sistem dengan keterbatasan sumber daya.

Analisis Berdasarkan Resolusi Citra

Hasil eksperimen juga menunjukkan bahwa peningkatan resolusi citra berdampak positif terhadap performa model. Baik CNN maupun ViT mengalami peningkatan akurasi seiring dengan meningkatnya resolusi dari 224×224 hingga 1024×1024. Namun, peningkatan performa pada ViT lebih signifikan dibandingkan CNN, yang menunjukkan bahwa ViT lebih mampu memanfaatkan informasi global pada citra resolusi tinggi.

Di sisi lain, peningkatan resolusi juga menyebabkan peningkatan waktu pelatihan secara signifikan, terutama pada ViT. Hal ini disebabkan oleh kompleksitas mekanisme self-attention yang meningkat seiring bertambahnya jumlah patch citra.



Gambar 2. Perbandingan Akurasi Model CNN dan ViT terhadap Resolusi Gambar

Pembahasan

Hasil penelitian ini menunjukkan bahwa Vision Transformer memiliki keunggulan dalam menangkap hubungan global antar bagian citra, sehingga mampu memberikan performa klasifikasi yang lebih tinggi dibandingkan CNN. Hal ini sejalan dengan karakteristik self-attention yang memungkinkan model memahami konteks citra secara menyeluruh.

Namun demikian, keunggulan tersebut datang dengan konsekuensi berupa peningkatan kompleksitas komputasi. CNN, dengan sifat konvolusinya yang lokal, tetap unggul dalam hal efisiensi waktu pelatihan dan inferensi. Oleh karena itu, dalam konteks aplikasi nyata, pemilihan model perlu mempertimbangkan kebutuhan spesifik, apakah lebih mengutamakan akurasi atau efisiensi.

Secara keseluruhan, penelitian ini menegaskan bahwa tidak terdapat satu model yang secara mutlak unggul dalam semua aspek. Vision Transformer lebih cocok digunakan pada aplikasi yang memerlukan akurasi tinggi dan mampu memanfaatkan sumber daya komputasi yang besar, sedangkan CNN tetap relevan untuk aplikasi yang membutuhkan efisiensi dan stabilitas.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa Vision Transformer (ViT) menunjukkan performa yang lebih unggul dibandingkan Convolutional Neural Network (CNN) dalam tugas klasifikasi citra resolusi tinggi. Hal ini ditunjukkan oleh nilai accuracy, precision, recall, dan F1-score yang secara konsisten lebih tinggi pada model ViT. Keunggulan ini menunjukkan bahwa mekanisme self-attention pada ViT lebih efektif dalam menangkap hubungan global antar bagian citra,

terutama pada citra dengan resolusi tinggi yang memiliki kompleksitas informasi yang besar.

Di sisi lain, CNN—khususnya arsitektur EfficientNet—masih menunjukkan keunggulan dari aspek efisiensi komputasi. Waktu pelatihan dan inferensi CNN lebih rendah dibandingkan ViT, sehingga lebih sesuai untuk implementasi pada lingkungan dengan keterbatasan sumber daya. Selain itu, peningkatan resolusi citra terbukti memberikan dampak positif terhadap performa kedua model, namun dengan konsekuensi peningkatan waktu komputasi yang lebih signifikan pada ViT.

Dengan demikian, penelitian ini menunjukkan adanya trade-off antara akurasi dan efisiensi komputasi dalam pemilihan model. Vision Transformer lebih direkomendasikan untuk aplikasi yang mengutamakan akurasi tinggi dan memiliki dukungan komputasi yang memadai, sedangkan CNN tetap menjadi pilihan yang relevan untuk aplikasi yang membutuhkan efisiensi dan stabilitas.

DAFTAR PUSTAKA

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. http://arxiv.org/abs/2010.11929

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , 770–778.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, Z., Hu, J., Wu, K., Miao, J., Zhao, Z., & Wu, J. (2024). Local feature acquisition and global context understanding network for very high-resolution land cover classification. *Scientific Reports*, *14*(1), 12597. https://doi.org/10.1038/s41598-024-63363-7

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences*, *13*(9), 5521. https://doi.org/10.3390/app13095521

Rawat, W., & Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, *29*(9), 2352–2449. https://doi.org/10.1162/neco_a_00990

Rodrigo, M., Cuevas, C., & García, N. (2024). Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks. *Scientific Reports*, *14*(1), 21392. https://doi.org/10.1038/s41598-024-72254-w

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., Machino, H., Kobayashi, K., Asada, K., Komatsu, M., Kaneko, S., Sugiyama, M., & Hamamoto, R. (2024). Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. *Journal of Medical Systems*, *48*(1), 84. https://doi.org/10.1007/s10916-024-02105-8

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning* , 6105–6114.

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

Wang, Y., Deng, Y., Zheng, Y., Chattopadhyay, P., & Wang, L. (2025). Vision Transformers for Image Classification: A Comparative Survey. *Technologies*, *13*(1), 32. https://doi.org/10.3390/technologies13010032

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)* , 13–19.