BITNET
Jurnal Pendidikan Teknologi Informasi
Vol. 11 No. 1, January 2026, pp. 73-86
https://journal.umpr.ac.id/index.php/bitnet/issue/...
DOI: https://doi.org/...

Analisis Komparatif Arsitektur CNN dan ViT dengan Variasi Augmentasi untuk Klasifikasi Karakter Mandarin

Stefanus Eko Prasetyo[1], Haeruddin[2], Wilsen Lau[3]
[1,2,3]Universitas Internasional Batam, Indonesia
*Corresponding Author: 2232064.wilsen@uib.edu

ABSTRACT

Handwritten Chinese Character Recognition (HCCR) is a fundamental task in computer vision that poses significant challenges due to the large number of character classes and high variability of writing styles. This study aims to perform a systematic comparative evaluation between Convolutional Neural Network (ResNet-50) and Vision Transformer (ViT-B/16) architectures in handling these challenges under limited data scenarios. The research utilizes a specific subset of the CASIA-HWDB1.1 dataset, consisting of 10 fine-grained character classes selected for their visual similarity. To ensure robust evaluation a 5-Fold Cross-Validation method was employed across five experimental scenarios (Baseline, Geometric, Elastic Transform, Random Erasing, and CutMix). Experimental results demonstrate that ResNet-50 consistently outperformed ViT-B/16 in terms of accuracy and stability. The highest performance was achieved by ResNet-50 using Elastic Transform augmentation with a test accuracy of 90.52%, whereas the best performance for ViT-B/16 was achieved using Random Erasing at 88.69%. The study also reveals that ViT exhibits higher performance variability compared to CNN. These findings conclude that for HCCR tasks with limited data, CNN possess a superior inductive bias for capturing local stroke features while data augmentation must be tailored specifically to the architecture type to maximize performance.

Keywords: Augmentation, HCCR, ViT, CNN, Cross-Validation, Fine-grained classification

Article history

Received:	*Revised:*	*Accepted:*	*Published:*
4 December 2025	*20 December 2025*	*5 January 2026*	*30 January 2026*

INTRODUCTION

Pengenalan Karakter Optik (OCR) telah menjadi teknologi yang matang, terutama untuk aksara Latin dan karakter cetak. Namun, pengenalan karakter Hanzi tulisan tangan (*Handwritten Chinese Character Recognition* - HCCR) secara offline tetap menyajikan tantangan yang kompleks dan berkelanjutan. Kesulitan ini bersumber dari beberapa karakteristik intrinsik aksara Hanzi yaitu jumlah kelas yang sangat besar, struktur goresan yang rumit dengan banyak kemiripan antar-kelas, serta variabilitas gaya penulisan yang ekstrem (Gan et al., 2020; Gui et al., 2023).

Untuk menjawab tantangan ini, pendekatan deep learning telah menjadi standar *de facto*. Dua paradigma arsitektur saat ini mendominasi bidang visi komputer: *Convolutional Neural Network* (CNN), yang diwakili oleh arsitektur matang seperti ResNet-50, dan *Vision Transformer* (ViT), yang diwakili oleh ViT-Base/16. ResNet-50 unggul dalam mengekstraksi fitur visual hierarkis, sementara ViT mengadopsi mekanisme self-attention untuk memodelkan hubungan kontekstual global di seluruh bagian citra (Dosovitskiy et al., 2021).

Convolutional Neural Network (CNN) telah menjadi 'tulang punggung' visi komputer. Filosofi dasar CNN terletak pada kemampuannya untuk mengeksploitasi lokalitas dalam gambar melalui operasi konvolusi. Keluarga arsitektur ResNet (*Residual Network*) merupakan salah satu puncak evolusi CNN, yang efektif untuk melatih jaringan yang sangat dalam dan sering digunakan sebagai baseline kuat dalam berbagai analisis perbandingan arsitektur (Dosovitskiy et

al., 2021; Yan & Huang, 2021). Keunggulan utama CNN terletak pada bias induktif yang dimilikinya membuat CNN sangat efisien dalam mempelajari pola visual bahkan dengan data yang Relatif terbatas. Meskipun demikian, penerapan arsitektur CNN yang sangat dalam seperti yang dirancang untuk *ImageNet* secara langsung pada citra karakter tidak selalu optimal karena ukuran citra yang lebih kecil dan jumlah data latih yang lebih sedikit (Meng et al., 2019).

Beranjak dari arsitektur konvensional, paradigma baru telah muncul: Vision Transformer (ViT). ViT menawarkan filosofi berbeda, memecah gambar menjadi serangkaian *patch* dan menggunakan mekanisme *self-attention* untuk menangkap dependensi global. Studi-studi aplikasi, seperti pada pengenalan objek umum (Dosovitskiy et al., 2021) maupun pada kasus spesifik seperti deteksi penggunaan masker (Jahja et al., 2023), secara konsisten melakukan evaluasi perbandingan untuk memvalidasi keunggulan dan karakteristik masing-masing arsitektur. Analisis mendalam mengenai cara ViT merespons berbagai jenis augmentasi juga telah menjadi topik penelitian tersendiri (Umakantha et al., 2021). Bahkan, beberapa studi aplikasi menunjukkan bahwa untuk dataset dengan jumlah data yang lebih sedikit, model seri CNN terbukti lebih cocok daripada arsitektur Vision Transformer (Ahn et al., n.d.). Lebih jauh lagi, performa model tidak hanya bergantung pada arsitekturnya, tetapi juga pada relevansi fitur yang dipelajari, di mana sebuah studi menunjukkan bahwa dataset pre-training dengan pengetahuan *domain-spesifik* (seperti struktur karakter) bisa lebih efektif daripada dataset umum berskala besar seperti *ImageNet* (Liu et al., 2022). Meskipun ViT mampu mencapai performa yang luar biasa pada dataset masif, arsitektur ini dikenal memiliki bias induktif yang rendah. Ini berarti ViT tidak memiliki asumsi bawaan tentang struktur lokal gambar seperti CNN, sehingga butuh data latih yang lebih besar untuk mempelajari pola tersebut.

Salah satu tantangan terbesar dalam melatih model *deep learning* yang tangguh adalah kebutuhan akan data latih dalam jumlah besar dan beragam (de Sousa Neto et al., 2024). Augmentasi data merujuk pada serangkaian teknik untuk memperbanyak sampel pelatihan secara artifisial, yang berfungsi sebagai bentuk regularisasi kuat untuk mencegah *overfitting* dan meningkatkan kemampuan generalisasi model (Jahja et al., 2023; Yan & Huang, 2021). Augmentasi data merujuk pada serangkaian teknik untuk memperbanyak sampel pelatihan secara artifisial dengan menciptakan versi modifikasi yang realistis dari data yang ada. Tujuannya adalah meningkatkan jumlah data untuk melatih model yang lebih dalam seperti ViT yang dikenal 'haus data', dan berfungsi sebagai bentuk regularisasi yang kuat untuk mencegah overfitting dan meningkatkan kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya (Jahja et al., 2023).

Efektivitas penerapan Deep Learning dan algoritma klasifikasi cerdas dalam menyelesaikan berbagai permasalahan pengenalan pola telah dikonfirmasi melalui sejumlah studi empiris terkini. (Haeruddin et al., 2023) membuktikan ketangguhan arsitektur CNN dalam menangani variasi fitur visual pada sistem pengenalan wajah (*facial recognition*), sementara (Sim & Yulianto, 2024) menekankan pentingnya evaluasi komparatif (benchmarking) yang ketat dalam mengukur kinerja model deteksi objek modern. Di ranah yang lebih luas, pendekatan klasifikasi berbasis *Machine Learning* juga terbukti adaptif untuk berbagai jenis data, sebagaimana dieksplorasi oleh (Christian et al., 2024). Meskipun metodologi-metodologi tersebut telah mapan pada domain aplikatif masing-masing, terdapat kebutuhan untuk menguji ketahanan arsitektur *Deep Learning* secara spesifik pada domain yang lebih menantang, yaitu pengenalan karakter tulisan tangan Mandarin yang memiliki karakteristik *fine-grained* dan kemiripan visual ekstrem.

Tantangan dalam pengenalan aksara Hanzi seringkali terletak pada pembedaan karakter fine-grained, yaitu karakter yang memiliki kemiripan visual sangat tinggi dan hanya dibedakan oleh detail goresan kecil atau rasio aspek. Penggunaan dataset berskala raksasa seringkali mengaburkan analisis mendalam mengenai perilaku spesifik model terhadap kasus-kasus sulit ini. Oleh karena itu, diperlukan penelitian yang terfokus pada subset karakter yang representatif untuk membedah secara mendalam bagaimana arsitektur model menangani fitur-fitur krusial tersebut tanpa bias dari dominasi kelas yang mudah dikenali.

Selain pemilihan arsitektur yang tepat, strategi augmentasi data memegang peranan vital dalam mengatasi risiko overfitting akibat keterbatasan data pada subset spesifik ini (Umakantha et al., 2021; Yan & Huang, 2021). Literatur menyajikan spektrum teknik yang luas, mulai dari

transformasi geometris, distorsi elastis, penghapusan acak (*random erasing*), dan *CutMix* (Nanni et al., 2022; Umakantha et al., 2021). Meskipun demikian, efektivitas berbagai strategi augmentasi ini dapat sangat bergantung pada karakteristik arsitektur model yang digunakan (Nanni et al., 2022; Yan & Huang, 2021). Penelitian terbaru oleh Galván et al (Galván et al., 2025) menegaskan bahwa evaluasi komparatif antara arsitektur konvensional (CNN) dan modern (Transformer) pada kondisi dataset spesifik Masih menjadi area riset yang sangat krusial untuk menentukan batas kemampuan masing-masing model.

Fokus penelitian utama saat ini meliputi evaluasi model pada subset karakter dengan kemiripan visual ekstrem, perbandingan *head-to-head* antara CNN dan ViT pada kondisi data terbatas untuk membuktikan teori efisiensi data, serta pengujian efektivitas jenis augmentasi spesifik terhadap arsitektur tertentu. Meskipun aspek-aspek tersebut telah mapan secara individual, terdapat celah penelitian signifikan pada titik temu ketiganya, di mana literatur yang ada cenderung berfokus secara terpisah atau hanya menguji varian model tanpa komparasi sistematis dalam konteks karakter Hanzi yang kompleks. Kesenjangan ini memunculkan pertanyaan fundamental yang belum terjawab, yakni apakah arsitektur ViT yang bersifat data-hungry mampu menyaingi efisiensi inductive bias CNN pada skenario dataset terbatas, apakah optimalisasi augmentasi bersifat universal atau justru spesifik terhadap karakteristik arsitektur, serta bagaimana kapabilitas kedua model dalam menangani tantangan klasifikasi fine-grained pada karakter dengan kemiripan tinggi.

METHOD

Penelitian ini menggunakan subset dari dataset standar CASIA-HWDB1.1. Subset dataset ini terdiri dari 10 kelas karakter tulisan tangan Mandarin sebagaimana disajikan dalam Tabel 1. Pemilihan ini dilakukan secara sengaja untuk memfokuskan analisis pada tantangan fine-grained classification, yaitu kemampuan model untuk membedakan karakter yang memiliki Tingkat kemiripan visual yang tinggi. Dataset dibagi menjadi dua bagian yaitu data latih dan data uji yang Terpisah sepenuhnya untuk menjaga objektivitas evaluasi.

Table 1. Daftar Karakter Penelitian

No	Karakter	Pinyin	Arti	Alasan Dipilih
1	人	rén	Orang	Pasangan mirip dengan '入'
2	入	rù	Masuk	Pasangan mirip dengan '人'
3	日	rì	Matahari	Pasangan mirip dengan '曰' (Beda rasio aspek)
4	曰	yuē	Berkata	Pasangan mirip dengan '日' (Beda rasio aspek)
5	己	jǐ	Diri sendiri	Pasangan mirip dengan '已' (Beda goresan kecil)
6	已	yǐ	Sudah	Pasangan mirip dengan '己' (Beda goresan kecil)
7	一	yī	Satu	Karakter sederhana (Kontrol)
8	十	shí	Sepuluh	Karakter sederhana (Kontrol)
9	我	wǒ	Saya	Karakter kompleks (Banyak goresan)
10	鷹	Yīng	Elang	Karakter kompleks (Banyak goresan)

Sebelum citra karakter dimasukkan ke dalam model, serangkaian tahapan pra-pemrosesan dilakukan untuk menjamin kualitas input serta kompatibilitas dengan arsitektur pre-trained. Proses ini diawali dengan konversi kanal, di mana citra asli dataset yang berformat grayscale (1 kanal) diubah menjadi format RGB (3 kanal) agar sesuai dengan spesifikasi input model ResNet-50 dan ViT-B/16. Selanjutnya, diterapkan teknik padding proporsional untuk mengakomodasi variasi rasio aspek pada karakter Mandarin. teknik ini menempatkan karakter di tengah kanvas guna menjaga proporsi asli dan mencegah distorsi bentuk saat pengubahan ukuran. Setelah itu, seluruh citra hasil padding dinormalisasi ukurannya (resizing) menjadi dimensi standar 224 x 224 piksel sesuai

resolusi input arsitektur. Tahapan akhir melibatkan normalisasi nilai piksel menggunakan parameter rata-rata dan standar deviasi dari ImageNet, yang bertujuan untuk mempercepat konvergensi selama proses pelatihan.

Penelitian ini melakukan evaluasi komparatif (*head-to-head*) terhadap dua arsitektur yang merepresentasikan paradigma dominan dalam computer vision saat ini, yaitu *Convolutional Neural Network* (CNN) dan *Vision Transformer* (ViT). Sebagai perwakilan arsitektur CNN, digunakan model ResNet-50 yang dipilih karena maturitas dan kinerjanya yang tinggi sebagai *baseline* standar dalam berbagai penelitian. mekanisme koneksi residual (*residual connections*) pada model ini secara efektif mengatasi masalah *vanishing gradient* dan memungkinkan pelatihan jaringan yang sangat dalam. Sementara itu, untuk mewakili paradigma *Vision Transformer*, penelitian ini menggunakan ViT-Base/16 yang memproses citra dengan membaginya menjadi serangkaian *patch* non-tumpang tindih berukuran 16 x 16 piksel ke dalam encoder Transformer standar. Pemilihan varian ViT-Base ini bertujuan untuk memastikan perbandingan yang adil dengan ResNet-50 serta mengevaluasi respons pendekatan arsitektur yang secara fundamental berbeda tersebut terhadap teknik augmentasi data yang sama.

Strategi inisialisasi model (*Transfer Learning*) mengingat keterbatasan jumlah sampel pada subset dataset yang digunakan, melatih model dari awal berisiko mengalami *overfitting* dan Konvergensi yang lambat. Literatur terbaru menunjukkan bahwa melatih ViT dari nol pada dataset kecil seringkali menghasilkan Performa suboptimal karena kurangnya bias induktif (Lee et al., 2022). Dengan demikian penelitian ini menerapkan strategi *Transfer Learning*. Kedua model diinisialisasi menggunakan bobot pre-trained yang sudah dilatih. Proses pelatihan dilakukan dengan metode *fine-tuning*, di mana bobot jaringan diizinkan untuk diperbarui dengan laju pembelajaran rendah, sementara lapisan klasifikasi terakhir diganti dan dilatih untuk memprediksi 10 kelas karakter target. Strategi ini terbukti efektif dalam berbagai studi untuk mempercepat konvergensi dan meningkatkan akurasi (Husen et al., 2023; Solak, 2024).

Untuk mengevaluasi pengaruh augmentasi data secara sistematis terhadap kedua arsitektur model (ResNet-50 dan ViT-Base/16), penelitian ini merancang lima skenario eksperimen yang berbeda. Skenario pertama berfungsi sebagai Baseline (*Group Kontrol*), di mana model dilatih menggunakan data original tanpa augmentasi sama sekali untuk menetapkan tolak ukur kinerja dasar (*baseline performance*). Skenario kedua menerapkan augmentasi geometris yang memanipulasi struktur spasial citra melalui rotasi acak dalam rentang kurang lebih 15 derajat dan penskalaan acak sebesar kurang lebih 10 derajat. Skenario ketiga berfokus pada teknik Distorsi Elastis, sebuah metode yang dirancang khusus untuk menguji efektivitas simulasi variasi non-linear alami pada goresan tulisan tangan manusia (de Sousa Neto et al., 2024). Selanjutnya, skenario keempat menguji teknik Random Erasing yang diperkenalkan oleh Zhong et al. (Zhong et al., 2017) sebagai metode regularisasi dengan menyembunyikan area persegi acak pada citra input untuk memaksa model mempelajari fitur secara holistik. Terakhir, skenario kelima menerapkan strategi CutMix (Baek et al., 2022), yang menghasilkan sampel pelatihan baru dengan memotong dan menempelkan patch antar-gambar serta mencampurkan label kelasnya secara proporsional.

Protokol pelatihan dan evaluasi dirancang secara ketat untuk menjamin validitas statistik dan optimalisasi model. Untuk memitigasi bias yang mungkin timbul dari pembagian data latih dan validasi yang statis, penelitian ini menerapkan metode *5-Fold Cross-Validation*. Dalam skema ini, dataset latih dibagi menjadi lima bagian (*folds*), dan eksperimen dilakukan sebanyak lima kali putaran. pada setiap iterasi, model dilatih

menggunakan empat bagian dan divalidasi pada satu bagian tersisa, dengan nilai performa akhir serta stabilitas dihitung berdasarkan rata-rata dari kelima putaran tersebut. Bersamaan dengan itu, untuk menemukan konfigurasi terbaik pada setiap skenario augmentasi, dilakukan pencarian hyperparameter otomatis (*Grid Search*). Parameter yang diuji meliputi *Learning Rate* (0.001 dan 0.0001), jenis *Optimizer* (AdamW dan SGD), serta ukuran *Batch* (32 dan 64, dengan batasan khusus 32 untuk ViT demi menjaga stabilitas memori), dengan durasi pelatihan dibatasi maksimal 15 *epoch*.

FINDINGS AND DISCUSSION

Findings

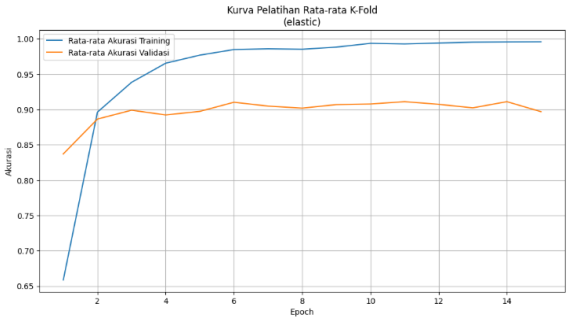
Hasil eksperimen perbandingan performa antara Convolutional Neural Network (ResNet-50) dan Vision Transformer (ViT-B/16) dalam hal klasifikasi karakter tulisan tangan mandarin diperoleh melalui proses validasi menggunakan *5-Fold Cross-Validation* pada dataset HWDB1.1 untuk menjamin stabilitas, validitas, dan reliabilitas hasil penelitian sebagai berikut:

1. Hasil Eksperimen Arsitektur ResNet-50

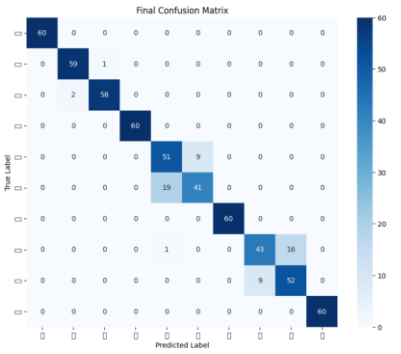
   Hasil Performa ResNet-50 pada lima scenario augmentasi data yang berbeda. Setiap skenario dievaluasi berdasarkan learning curve, confusion matrix, dan metrik Performa rata-rata.

   a. Skenario 1: *Elastic Transfrom* (Performa Terbaik)

   Skenario augmentasi *Elastic Transform* menghasilkan Performa tertinggi pada arsitektur ResNet-50 dengan akurasi mencapai 90.52%. Konfigurasi hyperparameter yang menghasilkan Performa terbaik dalam skenario ini dengan menggunakan learning rate sebesar 0.0001, batch size 32, dan optimizer AdamW dalam pelatihan maksimal 15 epoch. Pencapaian akurasi ini mengindikasikan bahwa penerapan distorsi elastis sangat efektif dalam mensimulasikan variasi goresan tangan manusia. Kemampuan augmentasi dalam memperkaya data terbukti berhasil membuat model menjadi lebih Tangguh saat menghadapi data test. Efisiensi proses pembelajaran terlihat dari rata-rata epoch terbaik yang dicapai pada posisi ke-8.6 yang menunjukkan bahwa model mencapai konvergensi dengan cukup cepat. Visualisasi tersebut, yang meliputi kurva pelatihan (*learning curve*) dan matriks konfusi (*confusion matrix*), disajikan secara lengkap pada Gambar 1 dan Gambar 2.



Gambar 1. Kurva Pelatihan

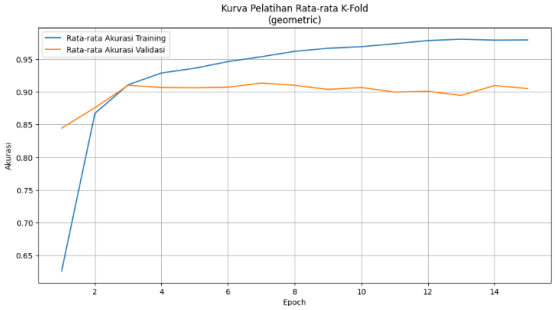


Gambar 2. Matrix Kebingungan

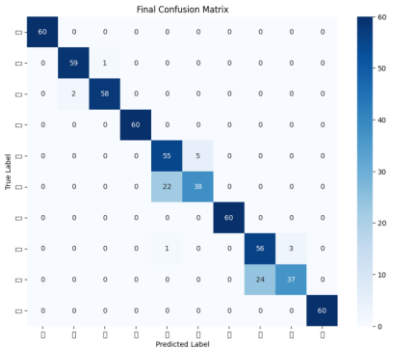
   b. Skenario 2: *Geometric Augmentation*

   Skenario ini menempati peringkat kedua dengan menghasilkan akurasi sebesar 90.35%. Model ini mencapai Konvergensi yang paling cepat di antara semua skenario pada epoch ke-6.2. Hal ini menunjukkan bahwa ResNet-50 memiliki

kemampuan adaptasi yang responsif terhadap variasi rotasi dan pergeseran pada gambar. Visualisasi tersebut, yang meliputi kurva pelatihan (*learning curve*) dan matriks konfusi (*confusion matrix*), disajikan secara lengkap pada Gambar 3 dan Gambar 4.



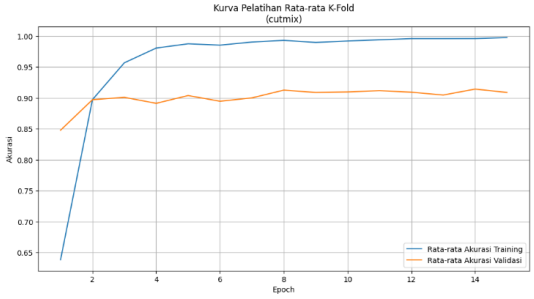
Gambar 3. Kurva Pelatihan



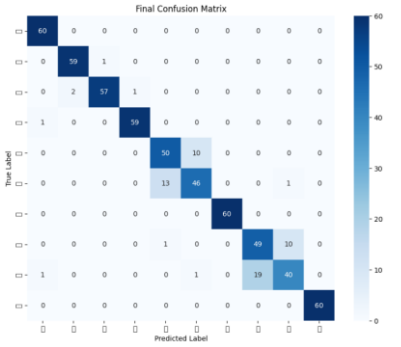
Gambar 4. Matrix Kebingungan

c. Skenario 3: *Cutmix Regularization*

Skenario ini menghasilkan stabilitas terbaik dengan standar deviasi terendah yaitu 0.0054 dan menghasilkan akurasi sebesar 89.85%. Meskipun akurasi sedikit dibawah *Elastic Transform*, Teknik *Cutmix* terbukti paling efektif dalam menjaga konsistensi Performa antar-fold. Namun model membutuhkan waktu yang lebih lama untuk mecapai Konvergensi. Epoch terbaik yang dicapai pada posisi ke-13.2 yang wajar mengingat kompleksitas akibat pemotongan dan pencampuran gambar. Visualisasi tersebut, yang meliputi kurva pelatihan (*learning curve*) dan matriks konfusi (*confusion matrix*), disajikan secara lengkap pada Gambar 5 dan Gambar 6.



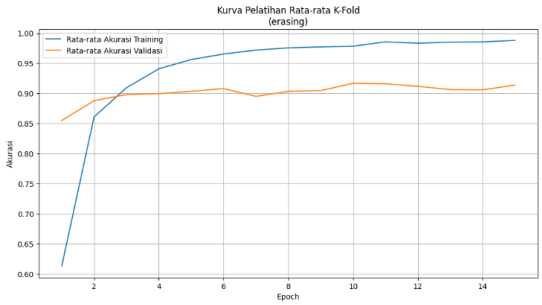
Gambar 5. Kurva Pelatihan

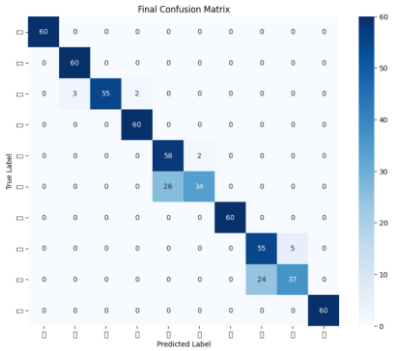


Gambar 6. Matrix Kebingungan

d. Skenario 4: *Random Erasing*

Skenario ini menghasilkan akurasi sebesar 89.68% dengan rata-rata akurasi validasi 92.39%. Visualisasi tersebut, yang meliputi kurva pelatihan (learning curve) dan matriks konfusi (confusion matrix), disajikan secara lengkap pada Gambar 7 dan Gambar 8.

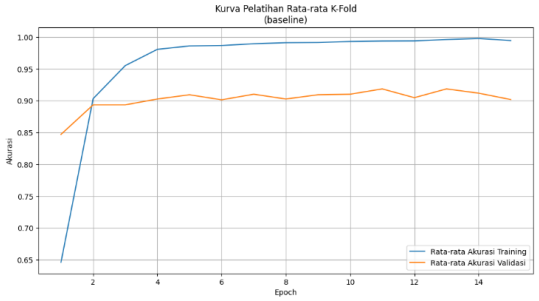


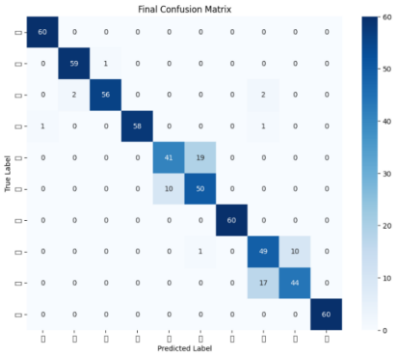


Gambar 7. Kurva Pelatihan **Gambar 8. Matrix Kebingungan**

e. Skenario 5: Baseline (Tanpa Augmentasi)

Sebagai kontrol, skenario tanpa augmentasi menghasilkan akurasi sebesar 89.35%. ini menunjukkan bahwa tanpa augmentasi pun ResNet-50 memiliki kemampuan ekstraksi fitur yang sangat baik. Namun, augmentasi seperti Elastic Transfrom mampu mendorong Performanya lebih jauh. Visualisasi tersebut, yang meliputi kurva pelatihan (learning curve) dan matriks konfusi (confusion matrix), disajikan secara lengkap pada Gambar 9 dan Gambar 10.





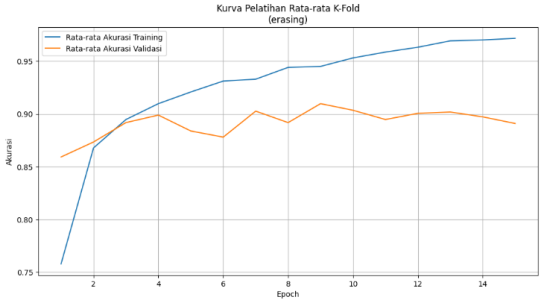
Gambar 9. Kurva Pelatihan **Gambar 10. Matrix Kebingungan**

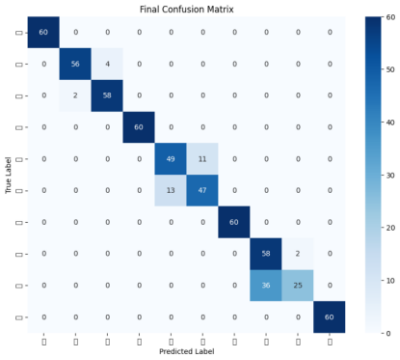
2. Hasil Eksperimen Arsitektur Vision Transformer (ViT)

Pada bagian ini memaparkan hasil Performa ViT-B/16 pada lima scenario augmentasi data yang berbeda. Setiap skenario dievaluasi berdasarkan *learning curve, confusion matrix*, dan metrik Performa rata-rata. Secara umum ViT menunjukkan variabilitas yang lebih tinggi dan akurasi absolut yang lebih rendah dibandingkan ResNet-50 pada dataset ini.

a. Skenario 1: Random Erasing (Performa Terbaik)

Berbeda dengan CNN, ViT mencapai performa tertingginya sebesar 88.69% menggunakan Teknik Random Erasing. Konfigurasi hyperparameter yang menghasilkan Performa terbaik dalam skenario ini dengan menggunakan learning rate sebesar 0.0001, batch size 32, dan optimizer AdamW dalam pelatihan maksimal 15 epoch. Ini menunjukkan kesesuaian yang kuat antara teknik Random Erasing dengan karakteristik Transformer yang mengandalkan mekanisme Self-Attention. Dengan menghapus sebagian area gambar secara acak, model dipaksa untuk tidak bergantung pada fitur lokal, melainkan dituntut untuk memahami hubungan global untuk mengompensasi informasi yang hilang. Pendekatan ini terbukti menjadi strategi pelatihan yang paling efektif untuk Vision Transformer karena mendorong model untuk membangun representasi karakter yang lebih utuh dan kontekstual. Visualisasi tersebut, yang meliputi kurva pelatihan (learning curve) dan matriks konfusi (confusion matrix), disajikan secara lengkap pada Gambar 11 dan Gambar 12.

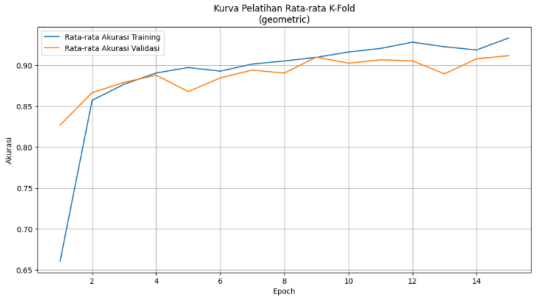


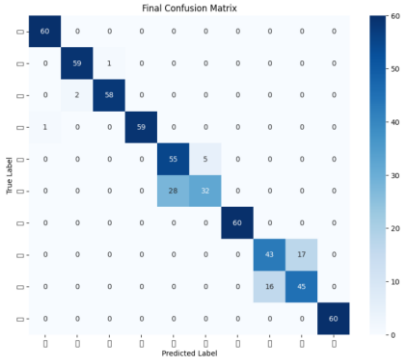


Gambar 11. Kurva Pelatihan **Gambar 12. Matrix Kebingungan**

b. Skenario 2: Geometric Augmentation

Skenario ini menghasilkan akurasi sebesar 88.35% dengan rata-rata validasi 91.93%. Visualisasi tersebut, yang meliputi kurva pelatihan (learning curve) dan matriks konfusi (confusion matrix), disajikan secara lengkap pada Gambar 13 dan Gambar 14.

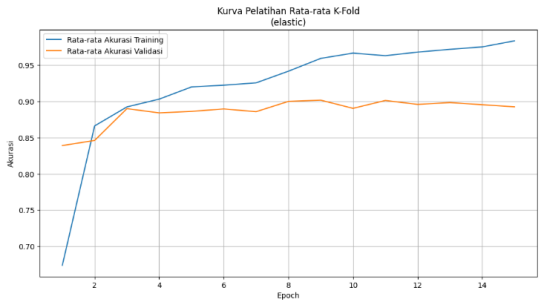


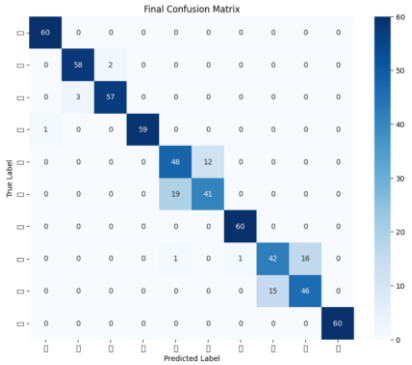


Gambar 13. Kurva Pelatihan **Gambar 14. Matrix Kebingungan**

c. Skenario 3: Elastic Transform

Skenario ini mencapai akurasi sebesar 88.35%, namun dengan stabilitas terendah. Hal ini menunjukkan bahwa ViT cukup sensitif terhadap distorsi elastis yang berlebihan yang menyebabkan performa yang fluktuatif antar-fold. Visualisasi tersebut, yang meliputi kurva pelatihan (learning curve) dan matriks konfusi (confusion matrix), disajikan secara lengkap pada Gambar 15 dan Gambar 16.

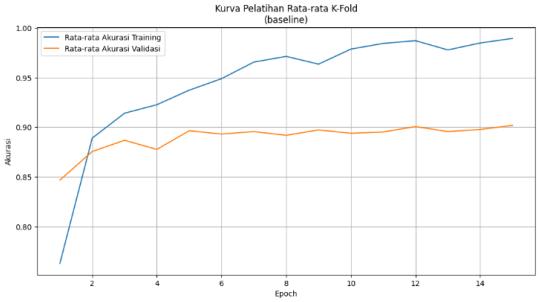


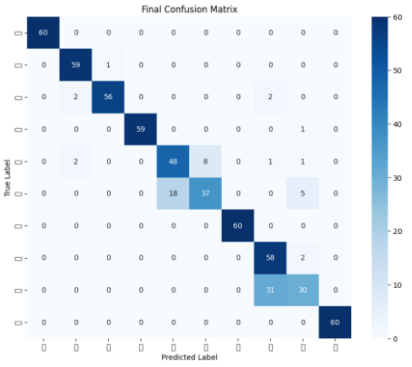


Gambar 15. Kurva Pelatihan **Gambar 16. Matrix Kebingungan**

d. Skenario 4: Baseline (Tanpa Augmentasi)

Pada skenario ini menghasilkan akurasi sebesar 87.69%. Visualisasi tersebut, yang meliputi kurva pelatihan (learning curve) dan matriks konfusi (confusion matrix), disajikan secara lengkap pada Gambar 17 dan Gambar 18.

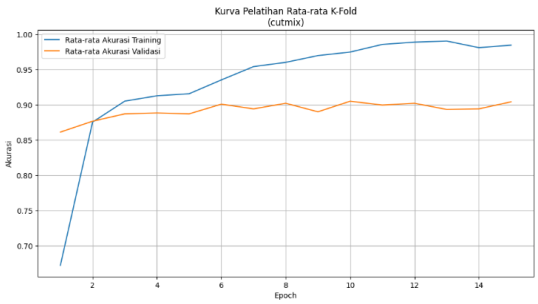




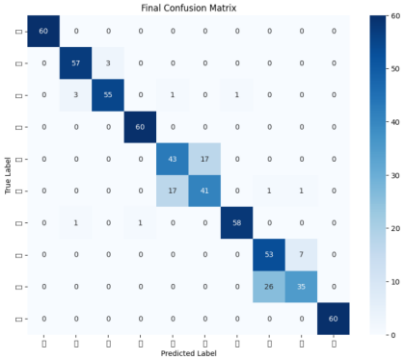
Gambar 17. Kurva Pelatihan **Gambar 18. Matrix Kebingungan**

e. Skenario 5: Cutmix

Pada skenario ini menghasilkan akurasi terendah untuk ViT sebesar 86.86%. Pencampuran gambar pada Teknik Cutmix merusak informasi posisi spasial dan struktur goresan yang krusial bagi ViT dalam mengenali karakter hanzi. Ini yang menyebabkan Performa yang signifikan dibandingkan skenario lainnya. Visualisasi tersebut, yang meliputi kurva pelatihan (learning curve) dan matriks konfusi (confusion matrix), disajikan secara lengkap pada Gambar 19 dan Gambar 20.



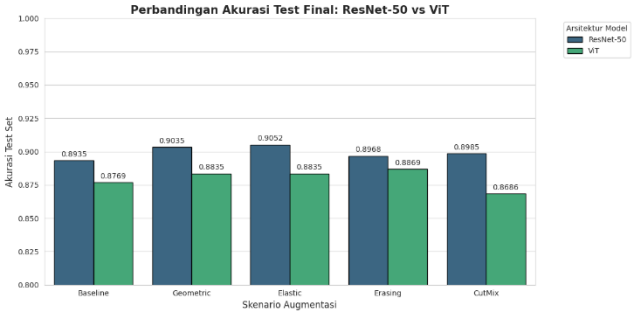
Gambar 19. Kurva Pelatihan



Gambar 20. Matrix Kebingungan

Discussion

Berdasarkan perbandingan head-to-head antara seluruh skenario augmentasi untuk kedua arsitektur. Visualisasi pada Gambar 21 dan Tabel 2 berikut merangkum peringkat Performa seluruh skenario eksperimen.



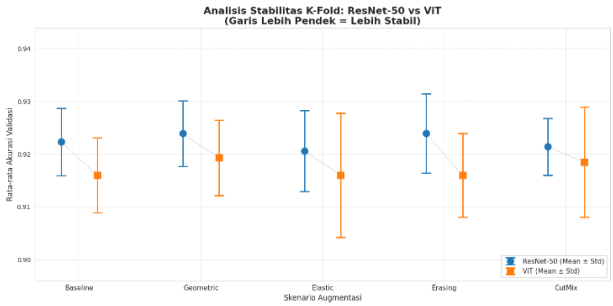
Gambar 21. Perbandingan Final

Table 2. Rangkuman Hasil Akhir

No	Model	Augmentasi	Akurasi Test Final	Rata-rata Validasi
1	ResNet-50	Elastic	0.9052	0.9206
2	ResNet-50	Geometric	0.9035	0.9239
3	ResNet-50	Cutmix	0.8985	0.9214
4	ResNet-50	Erasing	0.8968	0.9239
5	ResNet-50	Baseline	0.8935	0.9223
6	ViT	Erasing	0.8869	0.9160
7	ViT	Geometric	0.8835	0.9193
8	ViT	Elastic	0.8835	0.9160
9	ViT	Baseline	0.8769	0.9160
10	ViT	Cutmix	0.8686	0.9185

Dari data di atas jelas bahwa ResNet-50 mendominasi 5 peringkat teratas. Skenario terbaik ResNet-50 unggul sebesar 1.83% dibandingkan skenario terbaik ViT. Ini mengonfirmasi hipotesis bahwa untuk dataset dengan jumlah terbatas, arsitektur CNN lebih unggul dibandingkan *Vision Transformer*. Hal ini disebabkan oleh inductive bias pada CNN yaitu kemampuan untuk menangkap fitur visual lokal secara efisien tanpa memerlukan jumlah data yang Masif untuk pelatihan awal seperti yang dibutuhkan oleh ViT.

Stabilitas model selama validasi K-Fold dianalisis menggunakan plot standar deviasi. Visualisasi pada Gambar 22 menunjukkan bahwa ResNet-50 memiliki rentang error bar yang lebih pendek secara rata-rata dibandingkan ViT. Nilai standar deviasi ViT lebih tinggi dibandingkan ResNet-50 yang mengindikasikan bahwa Performa ViT lebih sensitive dan fluktuatif terhadap perubahan data latih dibandingkan ResNet yang lebih konsisten.



Gambar 22. Analisis Stabilitas K-Fold

Untuk memahami kesalahan kedua model, dilakukan analisis mendalam terhadap *Classification Report* dan *Confusion Matrix* pada kelas karakter yang memiliki tingkat kesalahan tertinggi. Berdasarkan analisis hasil prediksi, kedua model teridentifikasi mengalami kesulitan

signifikan pada dua pasang karakter yang memiliki tingkat kemiripan visual ekstrem. Kasus pertama ditemukan pada pasangan '己' (Jǐ) dan '已' (Yǐ) yang hanya dibedakan oleh ketinggian goresan vertikal terakhir. Pada kasus ini, ViT dengan skenario *Random Erasing* menunjukkan performa sedikit lebih unggul dengan F1-Score masing-masing 0.80 dan 0.79, dibandingkan ResNet-50 yang mencatat skor 0.77 dan 0.74. Keunggulan tipis ViT ini kemungkinan besar didorong oleh mekanisme *Global Attention* yang memungkinkannya melihat hubungan antar-goresan secara lebih utuh (*holistic*) dibandingkan CNN yang cenderung berfokus pada fitur lokal. Namun, dinamika berbeda terjadi pada kasus kedua, yaitu pasangan '日' (Rì) dan '曰' (Yuē) yang perbedaannya terletak pada rasio aspek (sempit/tinggi vs lebar/pendek). Pada pasangan ini, ViT mengalami kegagalan struktural yang serius. F1-Score terbaik untuk karakter '曰' hanya mencapai 0.56, bahkan recall-nya jatuh hingga 0.49 pada skenario Baseline. Sebaliknya, ResNet-50 mampu mempertahankan performa stabil dengan F1-Score 0.76 dan 0.80. Kegagalan ViT ini mengindikasikan kelemahan arsitektur tersebut dalam menangkap proporsi geometris pada dataset kecil, di mana mekanisme pemecahan citra menjadi patch berukuran 16 x 16 piksel diduga merusak informasi rasio aspek global yang menjadi fitur pembeda utama antara kedua karakter tersebut.

Pada kasus untuk karakter dengan bentuk yang unik dan memiliki kompleksitas goresan yang berbeda jauh dengan karakter lain seperti '我' (Wǒ - Saya), '鷹' (Yīng - Elang), dan '一' (Yī - Satu), kedua model mampu mencapai Performa sempurna yaitu mendapatkan F1-Score 1.00. Ini menunjukkan bahwa kedua model dapat melakukan klasifikasi karakter mandarin pada umumnya selama karakter tersebut memiliki fitur visual yang distingtif.

Hasil eksperimen ini sebagai studi perilaku arsitektur pada dataset terbatas. Temuan bahwa ResNet-50 (90.52%) mengungguli ViT-B/16 (88.69%) sejalan dengan temuan Sharma et al. (R et al., 2025), yang memberikan konfirmasi perbedaan cara belajar kedua arsitektur. CNN memiliki bias induktif kuat yang membuatnya efisien dan stabil pada sistem kritis (Filipiuk & Singh, 2019; Franche-comte & Franche-comte, n.d.). Sedangkan Vit bersifat data-hungry dan membutuhkan dataset masif untuk membentuk representasi fitur yang sekuat CNN, sebagaimana dijelaskan dalam analisis *robustness* oleh Bai et al. (Bai et al., 2021).

CONCLUSION

Berdasarkan eksperimen dan analisis komprehensif menggunakan *5-Fold Cross-Validation*, dapat disimpulkan bahwa arsitektur CNN secara konsisten mengungguli ViT pada dataset dengan jumlah kelas terbatas. ResNet-50 mencatat akurasi puncak sebesar 90.52%, mengungguli ViT-B/16 yang mencapai 88.69%. temuan ini mengonfirmasi bahwa CNN memiliki efisiensi data dan *inductive bias* yang lebih superior untuk menangkap fitur lokal pada skenario data non-masif.

penelitian ini juga mengungkapkan bahwa efektivitas teknik augmentasi sangat bergantung pada karakteristik arsitektur model. Teknik *Elastic Transform* terbukti paling optimal untuk ResNet-50 karena kemampuannya mensimulasikan variasi fisik natural penulisan tangan, sedangkan *Random Erasing* memberikan performa terbaik pada ViT dengan memaksa mekanisme *Self-Attention* mempelajari konteks global karakter secara utuh. Dari segi stabilitas, ResNet-50 menunjukkan konsistensi yang lebih tinggi, di mana teknik *CutMix* menghasilkan standar deviasi terendah (0.0054), berbeda dengan ViT yang menunjukkan fluktuasi performa lebih besar dan sensitivitas tinggi terhadap variasi data latih. Meskipun demikian, kedua model masih menghadapi tantangan signifikan dalam membedakan karakter dengan kemiripan visual ekstrem (*fine-grained*), khususnya pada pasangan '己' (Jǐ) vs '已' (Yǐ) yang dibedakan oleh detail goresan kecil, serta pasangan '日' (Rì) vs '曰' (Yuē) yang hanya berbeda pada rasio aspek.

Berdasarkan keterbatasan penelitian dan temuan empiris yang diperoleh, penulis mengajukan beberapa saran strategis untuk pengembangan penelitian selanjutnya. Pertama, mengingat karakteristik alami ViT yang membutuhkan data berskala masif untuk mencapai performa optimal, penelitian mendatang disarankan untuk memperluas cakupan evaluasi menggunakan dataset penuh CASIA-HWDB1.1 guna memvalidasi skalabilitas model. Kedua,

untuk mengurangi kelemahan intrinsik masing-masing arsitektur, disarankan untuk mengeksplorasi pengembangan arsitektur hibrida yang menggabungkan keunggulan ekstraksi fitur lokal CNN di tahap awal dengan kemampuan pemahaman konteks global ViT di tahap akhir. pendekatan ini berpotensi signifikan dalam meningkatkan akurasi pada karakter yang memiliki kemiripan visual tinggi. Terakhir, guna mengatasi masalah spesifik misklasifikasi pada pasangan karakter yang dibedakan oleh proporsi geometris (seperti '日' vs '曰'), penelitian masa depan dapat mengintegrasikan metode *aspect ratio preserving augmentation* atau menerapkan fungsi kerugian *Focal Loss* yang memberikan penalti lebih besar pada kesalahan klasifikasi sampel sulit (hard examples), sehingga model dapat belajar lebih fokus pada fitur pembeda yang halus.

REFERENCES

Ahn, J., Jang, T., Fengnyu, Q., Lee, H., Lee, J., & Lucia, S. (n.d.). *Enhancement of text recognition for hanja handwritten documents of Ancient Korea 1 Introduction*.

Baek, S., Park, J., Vepakomma, P., Raskar, R., Bennis, M., & Kim, S. (2022). *Visual Transformer Meets CutMix for Improved Accuracy, Communication Efficiency, and Data Privacy in Split Learning*.

Bai, Y., Mei, J., Yuille, A., & Xie, C. (2021). *Are Transformers More Robust Than CNNs ? NeurIPS*, 1–13.

Christian, Y., Wibowo, T., & Lyawati, M. (2024). *Sentiment Analysis by Using Naïve Bayes Classification and Support Vector Machine ,*. *8*(1), 258–275.

de Sousa Neto, A. F., Bezerra, B. L. D., de Moura, G. C. D., & Toselli, A. H. (2024). Data Augmentation for Offline Handwritten Text Recognition: A Systematic Literature Review. *SN Computer Science*, *5*(2), 1–20. https://doi.org/10.1007/s42979-023-02583-6

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. *ICLR 2021 - 9th International Conference on Learning Representations*.

Filipiuk, M., & Singh, V. (2019). *Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems*.

Franche-comte, B., & Franche-comte, B. (n.d.). *Convolutional Neural Network (CNN) vs Vision Transformer (ViT) for Digital Holography*.

Galván, A., Higuero, M., Atutxa, A., Jacob, E., & Saavedra, M. (2025). *Comparing CNN and ViT for Open-Set Face Recognition †*. 1–19.

Gan, J., Wang, W., & Lu, K. (2020). *Characters as Graphs: Recognizing Online Handwritten Chinese Characters via Spatial Graph Convolutional Network*. *1*. http://arxiv.org/abs/2004.09412

Gui, D., Chen, K., Ding, H., & Huo, Q. (2023). Zero-shot Generation of Training Data with Denoising Diffusion Probabilistic Model for Handwritten Chinese Character Recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14188 LNCS*, 348–365. https://doi.org/10.1007/978-3-031-41679-8_20

Haeruddin, Herman, & Hendri, P. P. (2023). *Jurnal Teknologi Terpadu PENGEMBANGAN APLIKASI EMOTION RECOGNITION DAN FACIAL RECOGNITION MENGGUNAKAN ALGORITMA LOCAL BINARY PATTERN*. *9*(1), 49–55.

Husen, N., Mulugeta, F., Habtamu, B., & Choe, S. (2023). *Vision-Transformer-Based Transfer Learning for Mammogram Classification*.

Jahja, H. D., Yudistira, N., & Sutrisno. (2023). Mask usage recognition using vision transformer with transfer learning and data augmentation. *Intelligent Systems with Applications*, *17*(November 2022), 200186. https://doi.org/10.1016/j.iswa.2023.200186

Lee, S., Lee, S., Song, B. C., & Member, S. (2022). Improving Vision Transformers to Learn Small-Size Dataset From Scratch. *IEEE Access*, *10*(September), 123212–123224. https://doi.org/10.1109/ACCESS.2022.3224044

Liu, L., Lin, K., Huang, S., Li, Z., Li, C., Cao, Y., & Zhou, Q. (2022). *Instance Segmentation for Chinese Character Stroke Extraction, Datasets and Benchmarks*. http://arxiv.org/abs/2210.13826

Meng, Y., Wu, W., Wang, F., Li, X., Nie, P., Yin, F., Li, M., Han, Q., Sun, X., & Li, J. (2019). Glyce: Glyph-vectors for Chinese character representations. *Advances in Neural Information Processing Systems*, *32*(NeurIPS), 1–12.

Nanni, L., Paci, M., Brahnam, S., & Lumini, A. (2022). Feature transforms for image data augmentation. *Neural Computing and Applications*, *34*(24), 22345–22356. https://doi.org/10.1007/s00521-022-07645-z

R, R. S., Sungheetha, A., Tiwari, M., & Pindoo, I. A. (2025). *Comparative Analysis of Vision Transformer and CNN Architectures in Medical Image Classification*. *Icsice 24*.

Sim, J. H., & Yulianto, A. (2024). *Evaluating YOLOv5 and YOLOv8 : Advancements in Human Detection*. *6*(4), 2999–3015. https://doi.org/10.51519/journalisi.v6i4.944

Solak, A. (2024). *A Comparative Analysis of Vision Transformers and Transfer Learning for Brain Tumor Classification*. *13*, 558–572. https://doi.org/10.29130/dubited.1521340

Umakantha, A., Semedo, J. D., Golestaneh, S. A., & Lin, W.-Y. S. (2021). *How to augment your ViTs? Consistency loss and StyleAug, a random style transfer augmentation*. http://arxiv.org/abs/2112.09260

Yan, E., & Huang, Y. (2021). Do CNNs Encode Data Augmentations? *Proceedings of the International Joint Conference on Neural Networks*, *2021-July*. https://doi.org/10.1109/IJCNN52387.2021.9534219

Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2017). *Random Erasing Data Augmentation*.