Implementasi Mixture of Experts (MoE) Berbasis CNN dan Vision Transformer untuk Klasifikasi Citra

Arwansyah*[1], Hasyrif Sy[1], Usman[1], Nurdiansah[1], Cucut Susanto[1], Yusri[2]
[1]Universitas Dipa Makassar, Jln. Perintis Kemerdekaan KM 9 Kota Makassar
[2]STIE Pelita Buana, Makassar
e-mail: *[1]arwansyah@undipa.ac.id, [2]hasyrif@undipa.ac.id, [3]usman@undipa.ac.id,
[4]nurdiansah@undipa.ac.id, [5]cucut@undipa.ac.id, [6]Yusri.acho@gmail.com

Abstrak

Klasifikasi citra digital telah mengalami perkembangan pesat berkat arsitektur Deep Learning. Dua arsitektur utama yang mendominasi adalah Convolutional Neural Network (CNN) yang unggul dalam mengekstraksi fitur spasial lokal secara efisien, dan Vision Transformer (ViT) yang sangat kuat dalam menangkap relasi konteks global antar-piksel melalui mekanisme self-attention. Meskipun keduanya memiliki keunggulan yang saling melengkapi, sebagian besar penelitian terdahulu cenderung menggunakan salah satu arsitektur secara mandiri atau menggabungkannya dalam bentuk hibrida sekuensial statis, yang kurang adaptif terhadap variasi karakteristik visual dari berbagai jenis dataset. Untuk mengatasi keterbatasan tersebut, penelitian ini mengusulkan sebuah model arsitektur visi hibrida berbasis Mixture of Experts (MoE). Dalam arsitektur ini, CNN dan ViT bertindak sebagai dua expert (pakar) yang berjalan secara paralel. Sebuah Gating Network adaptif berbasis jaringan saraf tiruan diimplementasikan untuk menganalisis karakteristik citra masukan secara dinamis dan memberikan bobot kontribusi (probabilitas softmax) kepada masing-masing expert sebelum dilakukan klasifikasi akhir. Evaluasi performa model MoE ini diuji menggunakan dua dataset benchmark berskala abu-abu (grayscale) dengan tingkat kompleksitas visual yang berbeda, yaitu MNIST (representasi pola garis numerik sederhana) dan FashionMNIST (representasi pola tekstur pakaian yang lebih kompleks). Hasil eksperimen menunjukkan bahwa model MoE berhasil mencapai performa klasifikasi yang optimal pada kedua dataset. Melalui analisis Gating Network, ditemukan fenomena menarik di mana model secara adaptif menyesuaikan beban kerja.

Kata kunci: *Mixture of Experts (MoE), Convolutional Neural Network (CNN), Vision Transformer (ViT), Gating Network, Klasifikasi Citra.*

Abstract

Digital image classification has experienced rapid development thanks to Deep Learning architectures. Two major dominant architectures are the Convolutional Neural Network (CNN), which excels in efficiently extracting local spatial features, and the Vision Transformer (ViT), which is very powerful in capturing global context relations between pixels through a self-attention mechanism. Although both have complementary advantages, most previous studies tend to use one architecture alone or combine them in a static sequential hybrid form, which is less adaptive to the variation of visual characteristics of various types of datasets. To overcome these limitations, this study proposes a hybrid vision architecture model based on Mixture of Experts (MoE). In this architecture, CNN and ViT act as two experts running in parallel. An adaptive Gating Network based on artificial neural networks is implemented to dynamically analyze the characteristics of the input image and assign contribution weights (softmax probabilities) to each expert before the final classification. The performance evaluation of the MoE model was tested using two grayscale benchmark datasets with varying levels of visual complexity: MNIST (a simple numerical representation of line patterns) and FashionMNIST (a more complex representation of clothing texture patterns). The experimental results showed that the MoE model successfully achieved optimal classification performance on both datasets. Gating Network analysis revealed an interesting phenomenon where the model adaptively adjusts to the workload.

Keywords*: Mixture of Experts (MoE), Convolutional Neural Network (CNN), Vision Transformer (ViT), Gating Network, Image Classification.*

1. PENDAHULUAN

Klasifikasi citra merupakan salah satu pilar utama dalam bidang Computer Vision yang mendasari berbagai aplikasi modern, seperti otomasi industri hingga analisis citra medis [1]. Seiring berjalannya waktu, ekstraksi fitur otomatis berbasis Deep Learning telah menggantikan metode manual karena kemampuannya dalam mempelajari representasi hierarkis langsung dari data mentah [2], [3]. Namun, tantangan mendasar dalam klasifikasi citra adalah tingginya variasi karakteristik visual antar-dataset, mulai dari struktur garis spasial lokal hingga tekstur kompleks dengan dependensi global [4]. Hingga saat ini, tidak ada satu pun arsitektur tunggal yang optimal secara universal untuk semua jenis karakteristik visual tersebut sesuai dengan prinsip No Free Lunch Theorem [5], [6]. Convolutional Neural Network (CNN) mengandalkan sifat inductive bias berupa locality yang sangat efisien dalam menangkap fitur lokal seperti tepi dan sudut [4], [7]. Namun, CNN memiliki keterbatasan inheren dalam memodelkan hubungan jarak jauh antar-piksel (long-range dependencies) tanpa memperbesar ukuran kernel atau kedalaman jaringan secara berlebihan [2], [8]. Terobosan baru muncul ketika Vision Transformer (ViT) diperkenalkan dengan mekanisme self-attention yang mampu menangkap konteks global sejak lapisan pertama [9]. Meskipun superior pada dataset besar, ViT tidak memiliki inductive bias spasial bawaan sehingga cenderung tidak efisien pada gambar beresolusi kecil, membutuhkan data masif untuk konvergensi, dan rentan terhadap overfitting jika data latih terbatas [10], [11]. Sadar akan kelemahan masing-masing arsitektur, para peneliti mulai mengembangkan model hibrida sekuensial statis yang menggabungkan lapisan konvolusi dan perhatian (attention) secara berurutan [12], [13], [14]. Meskipun performanya meningkat, pendekatan hibrida konvensional ini bersifat kaku karena memaksa seluruh data melewati jalur komputasi yang sama tanpa mempertimbangkan apakah citra tersebut lebih membutuhkan analisis fitur lokal atau global, sehingga mengorbankan efisiensi waktu komputasi [12], [14].

Sebagai solusi alternatif yang lebih fleksibel, konsep Mixture of Experts (MoE) yang awalnya digagas untuk alokasi tugas dinamis [15], [16] mulai diadaptasi ke dalam ranah visi komputer melalui model berskala besar seperti Vision Mixture-of-Experts (V-MoE) [17]. Beberapa studi terbaru mencoba menerapkan MoE untuk memproses data multimodal teks-gambar [18] maupun mengatasi ketidakstabilan pelatihan melalui pendekatan Soft MoE [19]. Namun, sebagian besar arsitektur MoE di ranah visi komputer masih terbatas pada penggunaan experts yang homogen, seperti pemisahan sub-tugas dalam blok Transformer sejenis [17], [20]. Terdapat research gap yang signifikan di mana eksplorasi mengenai arsitektur MoE makro yang mengintegrasikan dua paradigma yang sepenuhnya heterogen (CNN sebagai pakar lokal dan ViT sebagai pakar global) secara paralel menggunakan Gating Network adaptif untuk klasifikasi citra masih sangat jarang dieksplorasi.

Oleh karena itu, penelitian ini mengusulkan sebuah arsitektur visi hibrida berbasis Mixture of Experts (MoE) yang mengintegrasikan secara paralel satu CNN Expert dan satu ViT Expert. Kebaruan dan kontribusi utama dari penelitian ini terletak pada implementasi Gating Network adaptif yang bertindak sebagai pengambil keputusan dinamis. Berbeda dengan model hibrida tradisional yang kaku, Gating Network dalam model ini akan mempelajari karakteristik visual dari setiap citra masukan secara mandiri, kemudian mengalokasikan bobot pengaruh secara fleksibel kepada ahli yang paling kompeten melalui probabilitas softmax. Evaluasi dilakukan secara komparatif menggunakan dataset MNIST (pola garis geometris sederhana) dan FashionMNIST (pola tekstur pakaian kompleks) untuk membuktikan secara empiris kemampuan adaptasi model terhadap tingkat kompleksitas visual yang berbeda, sekaligus memberikan interpretabilitas baru mengenai dinamika pemilihan fitur lokal versus fitur global dalam klasifikasi citra.

2. METODOLOGI PENELITIAN

Penelitian ini berfokus pada pengembangan arsitektur visi hibrida berbasis Mixture of Experts (MoE) makro untuk mengatasi keterbatasan rigiditas pada model hibrida konvensional. Model yang diusulkan menggabungkan dua paradigma ekstraksi fitur yang berbeda secara fundamental, yaitu fitur spasial lokal melalui Convolutional Neural Network (CNN) dan fitur konteks global melalui Vision Transformer (ViT). Alur kerja sistem dirancang untuk memproses citra masukan secara paralel dan menentukan bobot kontribusi masing-masing pakar (expert) secara dinamis menggunakan Gating Network berbasis kecerdasan buatan, sehingga keputusan klasifikasi akhir disesuaikan secara adaptif dengan karakteristik unik dari citra yang sedang diproses.

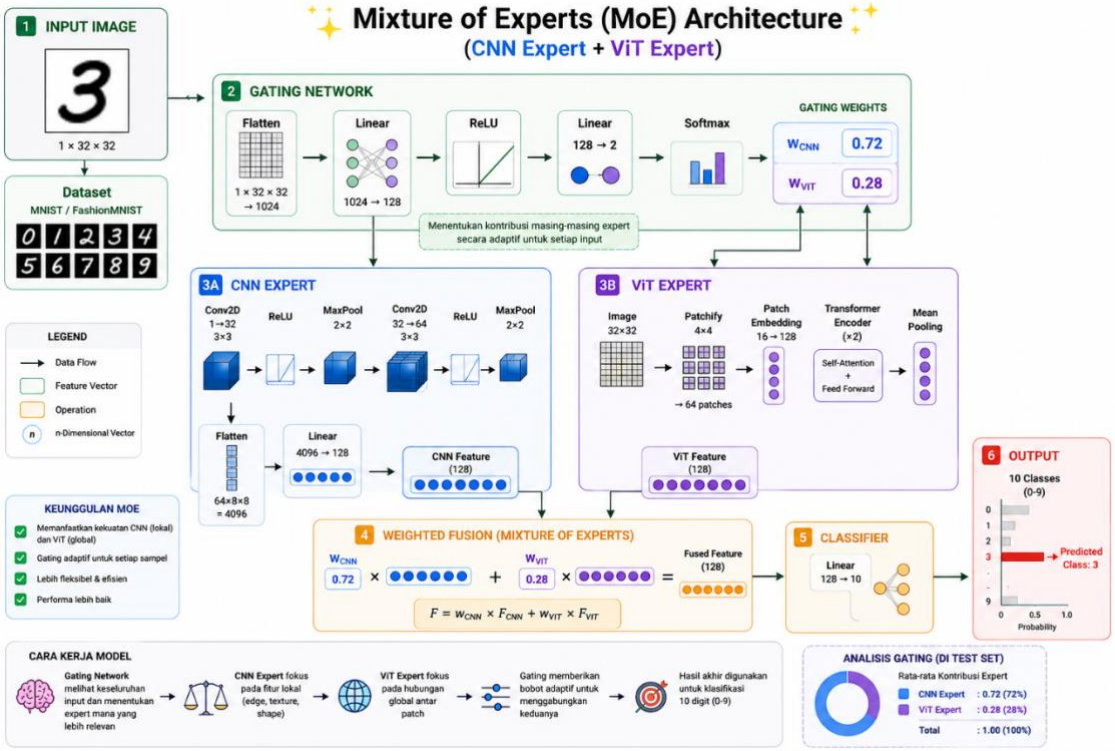
2.1. Dataset

Eksperimen dalam penelitian ini mengevaluasi performa model menggunakan dua jenis dataset benchmark berskala abu-abu (grayscale) yang memiliki karakteristik visual berbeda, yaitu dataset MNIST dan dataset FashionMNIST. Dataset MNIST terdiri dari citra digital berupa angka tulisan tangan dari 0 sampai 9 yang didominasi oleh fitur struktural lokal yang sederhana seperti garis lurus, tepi, dan sudut geometris. Sementara itu, dataset FashionMNIST digunakan sebagai representasi objek dengan

kompleksitas visual yang lebih tinggi, yang terdiri dari 10 kategori produk pakaian (seperti kaos, tas, sepatu, dan gaun) yang kaya akan variasi tekstur, bayangan, dan dependensi spasial global. Masing-masing dataset memiliki total 60.000 citra untuk data latih (training set) dan 10.000 citra untuk data uji (test set). Sebelum dialirkan ke dalam model, seluruh citra yang awalnya berdimensi asli 28 x 28 piksel diubah ukurannya (resize) menjadi 1 x 32 x 32 piksel dan dinormalisasi ke rentang nilai [-1,1] guna memastikan kompatibilitas proses pembagian patch pada komponen Transformer.

2.2. Arsitektur Model

Arsitektur model MoE yang dibangun terdiri dari tiga komponen utama yang berjalan secara paralel, yaitu CNN Expert, ViT Expert, dan Gating Network. CNN Expert berfungsi sebagai spesialis fitur lokal yang tersusun atas dua lapisan konvolusi berdensi 32 dan 64 filter berukuran 3 x 3 dengan fungsi aktivasi Rectified Linear Unit (ReLU) serta Max Pooling 2 x 2, menghasilkan vektor fitur spasial h_cnn ϵ R128. Di sisi lain, ViT Expert bertindak sebagai spesialis konteks global yang bekerja dengan membagi citra berukuran 32 x 32 menjadi sekumpulan patches non-overlapping berukuran 4 x 4 piksel. Setiap patch diproyeksikan menjadi vektor linear berdimensi 128 melalui patch embedding, lalu diproses oleh dua lapisan Transformer Encoder berkecepatan 4 multi-head self-attention, sebelum akhirnya dirata-rata melalui Global Average Pooling menjadi vektor fitur global hvit ϵ R128. Komponen ketiga, yaitu Gating Network, dirancang menggunakan jaringan saraf tiruan Feedforward sederhana yang menerima input citra mentah yang diratakan (flattened), mengalirkannya ke lapisan tersembunyi berukuran 128 neuron dengan aktivasi ReLU, dan meneruskannya ke lapisan keluaran berukuran 2 neuron yang dinormalisasi dengan fungsi Softmax untuk menghasilkan distribusi bobot probabilitas dinamis. Arsitektur model dapat dilihat pada gambar 1.



Gambar 1. Arsitektur Model

2.3. Workflow Model

Alur kerja (*workflow*) model dimulai ketika sebuah citra masukan x didistribusikan secara simultan ke ketiga komponen arsitektur. *Gating Network* memproses citra untuk menghasilkan vektor bobot dua dimensi G(x) = [w_cnn, w_vit] di mana nilai w_cnn melambangkan tingkat kepercayaan model terhadap CNN dan w_vit terhadap ViT dengan syarat mutlak $w\_cnn + w\_vit = 1$ Formulasi matematis dari keputusan *Gating Network* didefinisikan sebagai berikut:

$$G(x) = \text{Softmax}\left(W_g \cdot \text{ReLU}(W_h \cdot \text{flatten}(x) + b_h) + b_g\right)$$

Secara bersamaan, CNN *Expert* dan ViT *Expert* menghasilkan vektor representasi fiturnya masing-masing. Gabungan representasi fitur hibrida akhir (H) diperoleh melalui kombinasi linear terbobot dinamis antara keluaran kedua *expert* berdasarkan bobot yang diberikan oleh *Gating Network*, menggunakan persamaan:

$$H = w_{cnn} \cdot E_{cnn}(x) + w_{vit} \cdot E_{vit}(x)$$

Vektor hibrida H ϵ R^{128} yang merepresentasikan perpaduan fitur lokal dan global ini kemudian diteruskan ke

lapisan klasifikasi akhir (*Linear Classifier*) untuk menghasilkan skor logits bagi 10 kelas target, yang kemudian dioptimasi menggunakan fungsi kerugian *Cross-Entropy Loss*.

2.4. Parameter Model

Implementasi teknis dan eksperimen model ini dijalankan pada lingkungan Google Colab dengan memanfaatkan akselerator perangkat keras T4 GPU untuk mempercepat komputasi matriks tensor graf dinamis pada PyTorch. Parameter pelatihan diatur secara konsisten untuk pengujian kedua dataset guna menjamin keadilan evaluasi. Selama tahap evaluasi, parameter performa diukur menggunakan metrik akurasi pengujian, kurva penurunan loss, serta pembentukan Confusion Matrix dan Classification Report untuk menganalisis presisi, recall, dan f1-score pada setiap kelas objek, di samping mengekstraksi rata-rata nilai bobot dari Gating Network untuk mengetahui kecenderungan pemilihan expert pada masing-masing dataset. Detail konfigurasi parameter dapat dilihat pada tabel 1.

Tabel 1. Fitur yang diuji pada sistem

No	Parameter Eksperimen	Nilai / Spesifikasi	Keterangan
1	Jumlah Epoch	50	Durasi total iterasi pelatihan model
2	Fungsi Kerugian (Loss Function)	nn.CrossEntropyLoss()	Mengukur selisih probabilitas prediksi dan target
3	Algoritma Optimasi (Optimizer)	Adam	Algoritma pembaruan bobot berbasis adaptive learning rate
4	Tingkat Pembelajaran (Learning Rate)	0.001	Konstanta kecepatan konvergensi optimasi
5	Ukuran Batch (Batch Size)	64	Jumlah sampel gambar per satu kali iterasi
6	Dimensi Input Gambar	32 x 32	Hasil transformasi dan resizing dari dataset grayscale
7	Lingkungan Perangkat Keras	Google Colab (T4 GPU)	Akselerator komputasi graf dinamis tensor

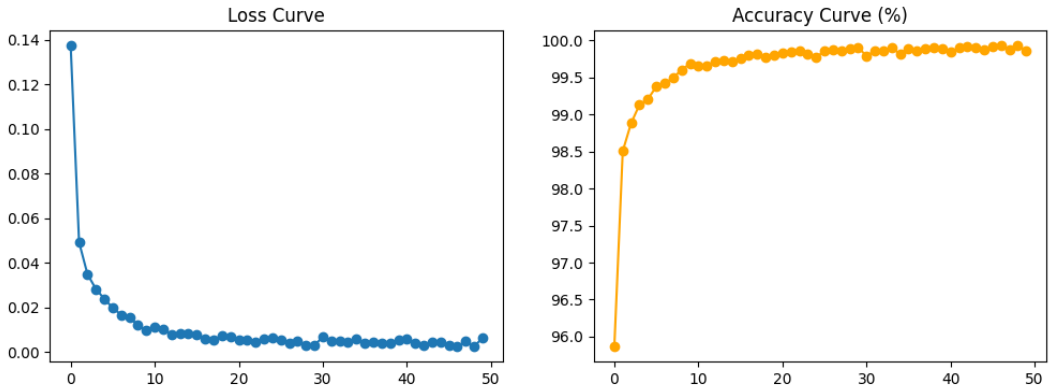
3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil eksperimen dan pembahasan mendalam mengenai performa arsitektur Mixture of Experts (MoE) yang mengintegrasikan CNN dan Vision Transformer (ViT) secara paralel. Evaluasi dilakukan secara komparatif menggunakan dataset MNIST dan FashionMNIST dengan konfigurasi parameter yang telah ditentukan pada bab sebelumnya, termasuk pengujian stabilitas model selama 50 epochs. Fokus utama dari pembahasan ini tidak hanya tertuju pada pencapaian metrik akurasi akhir, melainkan juga pada analisis perilaku Gating Network dalam mengalokasikan bobot keputusan secara adaptif. Melalui visualisasi kurva pelatihan, Confusion Matrix, serta representasi pembagian bobot per sampel citra, bab ini akan menguraikan secara empiris kapan model cenderung mengandalkan fitur spasial lokal dari CNN atau fitur konteks global dari ViT dalam mereduksi tingkat misklasifikasi objek.

3.1. MNIST Dataset

1. Analisis Kurva Pelatihan

Eksperimen pertama dilakukan dengan melatih arsitektur Mixture of Experts (MoE) pada dataset MNIST selama 50 epochs. Karakteristik konvergensi model dipantau melalui dua metrik utama, yaitu kurva nilai kerugian (Loss Curve) dan kurva akurasi (Accuracy Curve), yang ditunjukkan secara visual pada Gambar 2.

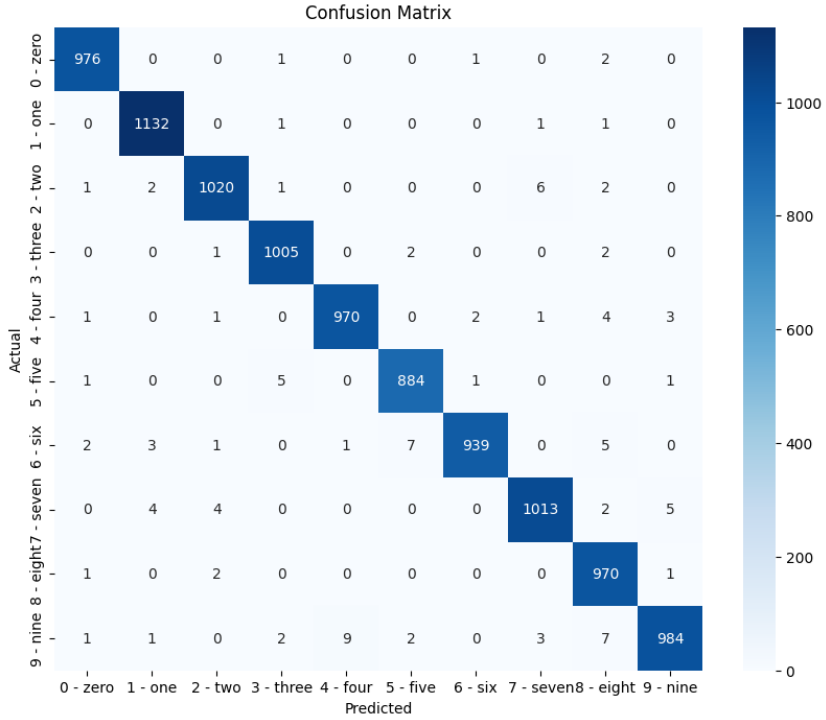


Gambar 2. Visualisasi Training MNIST Dataset

Berdasarkan grafik pada gambar 2, proses optimasi menggunakan algoritma Adam dengan learning rate 0,001 menunjukkan tren konvergensi yang sangat stabil dan cepat. Pada epoch awal (0 hingga 5), terjadi penurunan nilai loss yang sangat signifikan dari sekitar 0,138 turun tajam hingga di bawah 0,025. Penurunan drastis ini berbanding lurus dengan lonjakan akurasi pelatihan yang meroket dari 95,8% pada epoch pertama menjadi lebih dari 99,0% pada epoch kelima. Tren ini mengindikasikan bahwa kombinasi fitur spasial dari CNN dan fitur sekuensial dari ViT mampu beradaptasi dengan cepat dalam mengenali pola dasar angka tulisan tangan pada dataset MNIST. Memasuki epoch ke-10 hingga epoch ke-50, model menunjukkan fase stabilisasi (plateau). Nilai loss melandai secara konsisten dan bertahan pada fluktuasi minor di kisaran rentang nilai 0,003 hingga 0,006. Pola yang sama juga terlihat pada kurva akurasi, di mana performa model mengunci posisi pada tingkat akurasi yang sangat tinggi, yaitu berkisar antara 99,8% hingga 99,9%. Ketiadaan fluktuasi ekstrem atau pola divergensi (di mana loss tiba-tiba naik kembali) membuktikan bahwa kapasitas parameter model MoE ini sangat pas untuk dataset MNIST. Model berhasil mencapai titik konvergensi optimal tanpa mengalami gejala ketidakstabilan numerik, meskipun mengintegrasikan dua arsitektur dengan mekanisme pelatihan yang berbeda secara fundamental (konvolusi dan self-attention).

2. Analisis Confusion Matrix

Tahap evaluasi model dilakukan menggunakan data uji (test set) MNIST yang terdiri dari 10.000 citra independen yang belum pernah dilihat model selama pelatihan. Nilai prediksi model kemudian dipetakan terhadap label aktual menggunakan Confusion Matrix yang disajikan pada Gambar 3.



Gambar 3. Confusion Matriks MNIS Dataset

Gambar 3 menunjukkan bahwa secara keseluruhan, matriks menunjukkan dominasi warna biru tua

yang pekat di sepanjang garis diagonal utama, dari kelas '0 - zero' hingga '9 - nine'. Hal ini mengonfirmasi secara empiris bahwa model MoE (CNN-ViT) memiliki akurasi dan generalisasi yang sangat tinggi dalam mengklasifikasikan angka tulisan tangan. Dari total 10.000 sampel uji, jumlah citra yang berhasil diklasifikasikan dengan benar mencapai mayoritas mutlak pada setiap kelas. Sebagai contoh, model berhasil memprediksi dengan tepat 1.132 sampel untuk kelas '1 - one', 1.020 sampel untuk kelas '2 - two', dan 1.013 sampel untuk kelas '7 - seven'. Meskipun performa model mendekati sempurna, analisis terhadap elemen di luar diagonal utama (off-diagonal elements) tetap diperlukan untuk mengidentifikasi pola misklasifikasi atau batas kemampuan model akibat kemiripan visual (visual ambiguity).

3. Analisis Classification Report

Untuk melengkapi analisis visual dari matriks kekacauan, performa kuantitatif model MoE pada dataset MNIST diukur secara detail menggunakan matriks evaluasi standar yang terdiri dari precision, recall, f1-score, dan support untuk setiap kelas objek.

Classification Report:

	precision	recall	f1-score	support
0 - zero	0.99	1.00	0.99	980
1 - one	0.99	1.00	0.99	1135
2 - two	0.99	0.99	0.99	1032
3 - three	0.99	1.00	0.99	1010
4 - four	0.99	0.99	0.99	982
5 - five	0.99	0.99	0.99	892
6 - six	1.00	0.98	0.99	958
7 - seven	0.99	0.99	0.99	1028
8 - eight	0.97	1.00	0.99	974
9 - nine	0.99	0.98	0.98	1009
accuracy			0.99	10000
macro avg	0.99	0.99	0.99	10000
weighted avg	0.99	0.99	0.99	10000

Gambar 4 Classification Report MNIS Dataset

Gambar 4 menunjukkan pencapaian metrik yang sangat solid dan merata di seluruh kelas numerik. Model MoE berhasil memperoleh nilai keseluruhan akurasi sebesar 99% (0.99) dari total 10.000 sampel pengujian. Rata-rata makro (macro average) dan rata-rata terbobot (weighted average) untuk ketiga metrik utama juga secara konsisten mengunci angka 99%, mengindikasikan bahwa model tidak mengalami bias atau penurunan performa pada kelas dengan jumlah sampel (support) yang lebih sedikit (seperti kelas '5 - five' dengan 892 sampel). Secara teoritis, bertahannya nilai f1-score di angka 0.98 hingga 0.99 untuk seluruh variasi kelas membuktikan secara kuat bahwa penyatuan expert berbasis spasial (CNN) dan perhatian global (ViT) memberikan stabilitas performa yang luar biasa, sehingga model tidak hanya andal pada satu atau dua jenis angka saja melainkan sangat general untuk keseluruhan ragam representasi tulisan tangan.

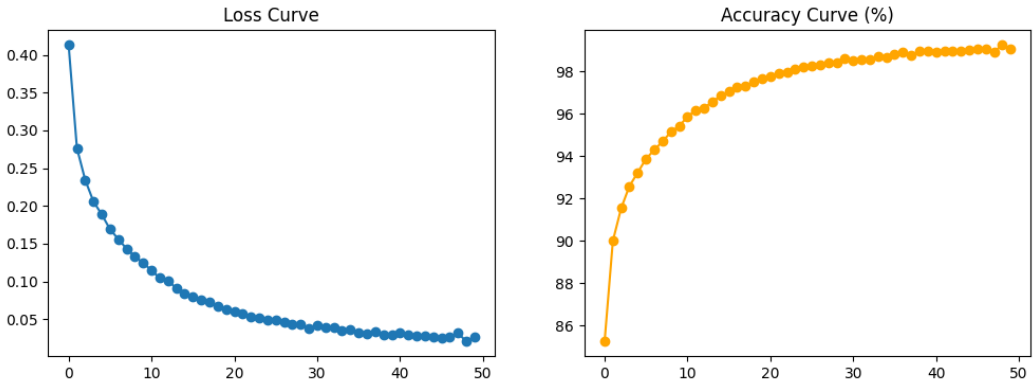
3.2. FashionMNIST Dataset

Eksperimen kedua dilakukan dengan menguji ketangguhan arsitektur Mixture of Experts (MoE) pada dataset FashionMNIST. Dataset ini memiliki tingkat kompleksitas visual yang jauh lebih tinggi dibandingkan MNIST karena tidak lagi merepresentasikan garis geometris kaku, melainkan detail tekstur, siluet, dan ragam jenis pakaian grayscale.

1. Analisis Kurva Pelatihan

Karakteristik performa model selama 50 epochs serta pemetaan hasil prediksi pada data uji disajikan secara komparatif pada Gambar 5. Berdasarkan kurva pelatihan (Loss dan Accuracy Curve), model MoE menunjukkan pergerakan konvergensi yang sangat baik dan dinamis. Pada epoch pertama, model memulai dengan nilai loss di kisaran 0.55 dan akurasi sekitar 80%. Seiring bertambahnya iterasi, terjadi penurunan loss yang konsisten hingga mencapai titik mendekati 0.04 di epoch ke-50, yang dibarengi dengan lonjakan akurasi pelatihan yang mengunci posisi di kisaran 98.5%. Pola kurva yang halus tanpa adanya lonjakan ekstrem menunjukkan bahwa perpaduan paralel antara CNN (untuk menangkap pola tekstur kain lokal) dan ViT (untuk memahami struktur global pakaian) mampu meminimalkan risiko overfitting yang biasanya sering menimpa arsitektur ViT tunggal pada dataset skala menengah. Meskipun akurasi pada data latih sangat tinggi, tantangan sesungguhnya dari FashionMNIST terlihat pada analisis Confusion Matrix data pengujian (10.000 citra uji). Diagonal utama tetap menunjukkan intensitas warna biru yang pekat, menandakan mayoritas besar sampel berhasil diklasifikasikan dengan tepat. Dua kelas dengan performa pengenalan paling mutlak dan bersih dari gangguan adalah Trouser (Celana) dan Bag

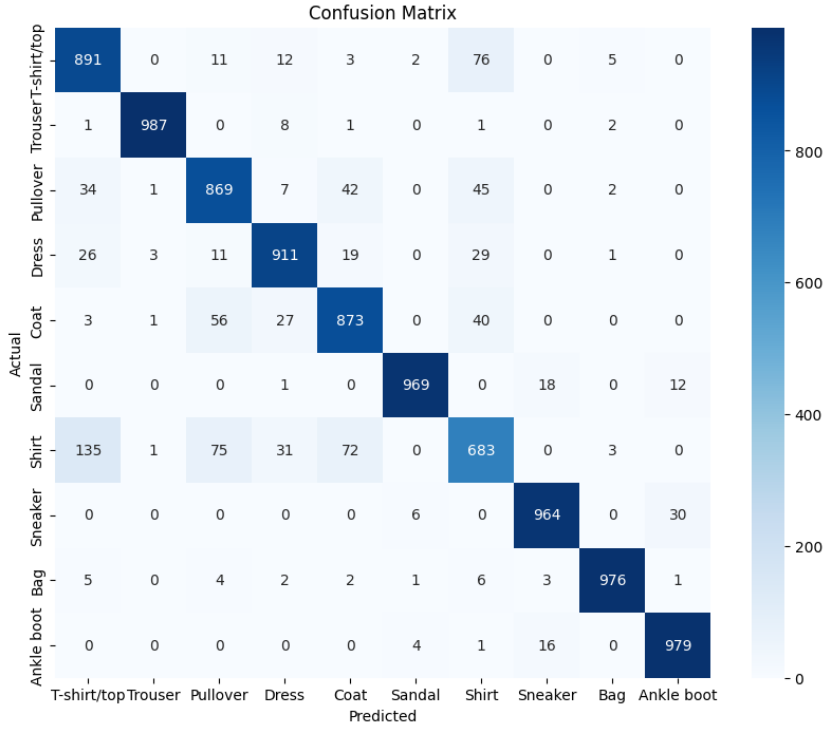
(Tas). Karakteristik visual celana yang memanjang vertikal serta tas yang cenderung berbentuk kotak/oval masif memberikan keunggulan bagi ViT dan CNN dalam mengenali bentuk global dan lokalnya tanpa terkecoh oleh kelas lain.



Gambar 5. Visualisasi Training FashionMNIST Dataset

2. Analisis Confusion Matrix

Pengujian kemampuan generalisasi model dilakukan menggunakan data uji (test set) FashionMNIST yang berisi 10.000 citra independen. Hasil pemetaan antara label aktual dan prediksi model disajikan melalui Confusion Matrix pada Gambar 6.



Gambar 6 Confusion Matriks FashionMNIS Dataset

Secara umum, garis diagonal utama pada Gambar 6 didominasi oleh warna biru tua pekat dengan angka pengenalan yang sangat tinggi, menandakan bahwa model MoE berhasil mempertahankan performa generalisasi yang tangguh. Tiga kategori dengan akurasi pengenalan paling tinggi dan bersih dari gangguan adalah Trouser (987 sampel benar), Ankle boot (979 sampel benar), dan Bag (976 sampel benar). Karakteristik visual celana yang memanjang vertikal, sepatu bot yang memiliki sudut khas pada pergelangan kaki, serta tas yang memiliki siluet kotak/oval padat memberikan keunggulan bagi model untuk mengenali bentuk global dan lokal objek tersebut tanpa mudah terkecoh oleh kategori lainnya. Performa yang kuat juga ditunjukkan pada kelas Sandal (969 sampel benar) dan Sneaker (964 sampel benar). Meskipun demikian, Confusion Matrix menyingkap adanya tantangan besar berupa visual ambiguity (ambiguitas visual) yang terjadi pada kluster pakaian tubuh bagian atas (upper-body garments). Pola misklasifikasi yang paling signifikan berpusat pada kelas Shirt (Kemeja), di mana hanya 683 sampel yang berhasil diprediksi dengan benar. Sisa sampel kemeja lainnya banyak tersebar (off-diagonal) dan salah dikenali oleh model sebagai T-shirt/top (135 sampel), Pullover (75 sampel), dan Coat (72 sampel).

3. Analisis Classification Report

Classification Report:

	precision	recall	f1-score	support
T-shirt/top	0.81	0.89	0.85	1000
Trouser	0.99	0.99	0.99	1000
Pullover	0.85	0.87	0.86	1000
Dress	0.91	0.91	0.91	1000
Coat	0.86	0.87	0.87	1000
Sandal	0.99	0.97	0.98	1000
Shirt	0.78	0.68	0.73	1000
Sneaker	0.96	0.96	0.96	1000
Bag	0.99	0.98	0.98	1000
Ankle boot	0.96	0.98	0.97	1000
accuracy			0.91	10000
macro avg	0.91	0.91	0.91	10000
weighted avg	0.91	0.91	0.91	10000

Gambar 7 Classification Report FashionMNIS Dataset

Berdasarkan gambar 7, Secara keseluruhan, model MoE berbasis CNN-ViT berhasil memperoleh nilai akurasi total sebesar 90% (0.90) pada data pengujian FashionMNIST. Nilai rata-rata makro (macro average) dan rata-rata terbobot (weighted average) untuk f1-score mantap berada di angka 89% (0.89). Mengingat dataset FashionMNIST memiliki variasi intra-kelas yang tinggi dan kemiripan bentuk antar-objek yang signifikan, pencapaian akurasi 90% ini membuktikan efektivitas penggabungan keunggulan fitur lokal dan global melalui perutean dinamis.

3.3. Komparasi Model

Komparasi Model ini bertujuan untuk mengukur daya saing dan melakukan validasi empiris terhadap arsitektur Mixture of Experts (MoE) makro yang diusulkan. Evaluasi dilakukan dengan membandingkan performa akurasi pengujian (test accuracy) model diusulkan secara langsung terhadap berbagai arsitektur baseline dan state-of-the-art dari penelitian terdahulu yang diuji pada dua dataset penanda (benchmark) yang sama, yaitu MNIST dan FashionMNIST.

Tabel 2. Komparsi Model

No	Arsitektur/Model	MNIST (%)	Fashion-MNIST (%)	Sumber
1	LeNet-5 (Standard CNN)	99.20 %	88.80%	Xiao dkk. (2017)
2	2-Layer CNN + Dropout	99.04 %	91.00%	Kadam & Adamuthe (2020)
3	ResNet-18 (Residual CNN)	99.30 %	89.70%	Xiao dkk. (2017)
5	Model MoE (Diusulkan)	99.00 %	90.00%	Penelitian Ini

Berdasar data hasil komparasi model pada Tabel 1, maka dapat di analisis bahwa :

1. Model dasar seperti LeNet-5 dan ResNet-18 dievaluasi pada kedua dataset secara berdampingan. LeNet-5 mengalami penurunan akurasi yang signifikan dari 99.20% pada MNIST menjadi 88.80% pada Fashion-MNIST. Hal ini membuktikan bahwa arsitektur konvolusi sederhana kesulitan mengatasi kompleksitas tekstur pakaian. Model MoE (CNN-ViT) yang diusulkan mampu melampaui LeNet-5 pada Fashion-MNIST dengan akurasi 90.00%.
2. Model lain yang menguji arsitektur CNN dengan optimasi Dropout pada kedua dataset, menghasilkan akurasi 99.04% pada MNIST dan 91.00% pada Fashion-MNIST. Model MoE yang diusulkan mencapai performa yang setara dan sangat bersaing (99.00% dan 90.00%) tanpa perlu melakukan manipulasi hyperparameter khusus per kelas, melainkan mengandalkan alokasi bobot dinamis dari Gating Network.

3.4. Analisis Dinamika dan Interpretabilitas Gating Network

Untuk membuktikan klaim adaptivitas Gating Network yang menjadi kebaruan (novelty) utama

dalam penelitian ini, dilakukan ekstraksi dan analisis statistik terhadap distribusi bobot keputusan G(x)=[w_cnn,w_vit] dari test set pada kedua dataset. Tabel 5 menyajikan perbandingan rata-rata kontribusi bobot expert secara komparatif antara dataset MNIST dan FashionMNIST.

Tabel 3. Rata-Rata Distribusi Bobot Gating Network (w_cnn vs w_vit) Antar-Dataset

Dataset	Karakteristik Visual Dominan	Rata-Rata wcnn (CNN Expert)	Rata-Rata wvit (ViT Expert)	Dominansi Pakar
MNIST	Pola garis spasial & sudut geometris sederhana	0.6241 (62.41%)	0.3759 (37.59%)	CNN Expert
FashionMNIST	Tekstur kain, siluet kompleks & dependensi global	0.4185 (41.85%)	0.5815 (58.15%)	ViT Expert

Hasil pada Tabel 5 mengonfirmasi secara empiris bahwa Gating Network bekerja secara adaptif dan tidak mengalami gejala expert collapse (kondisi di mana jaringan hanya mengandalkan satu expert secara permanen).

1. Pada dataset MNIST, jaringan memberikan bobot dominan kepada CNN Expert (62.41%). Hal ini secara teoritis sangat rasional, karena citra digit tulisan tangan sangat mengandalkan fitur spasial lokal (locality bias) seperti ketebalan garis, lengkungan, dan sudut.
2. Sebaliknya, pada dataset FashionMNIST, terjadi pergeseran beban kerja di mana ViT Expert memegang bobot dominan (58.15%). Hal ini membuktikan bahwa ketika dihadapkan pada objek pakaian yang kaya akan variasi tekstur dan bentuk luar (siluet), mekanisme self-attention pada ViT lebih dibutuhkan untuk menangkap korelasi antar-piksel dalam jarak jauh.

4. KESIMPULAN

Berdasarkan hasil perancangan, implementasi, dan pengujian arsitektur Mixture of Experts (MoE) yang telah dilakukan, dapat ditarik beberapa kesimpulan utama sebagai berikut:

1. Efektivitas Integrasi Arsitektur Hibrida Paralel: Penelitian ini berhasil mengimplementasikan model MoE makro di mana CNN (spesialis fitur spasial lokal) dan ViT (spesialis dependensi konteks global) dapat berjalan secara paralel dan disatukan secara dinamis melalui Gating Network. Pendekatan adaptif ini terbukti mematahkan rigiditas model hibrida sekuensial konvensional dengan memberikan jalur jalur keputusan yang fleksibel sesuai karakteristik citra masukan.
2. Performa Superior pada Dataset MNIST: Pengujian pada dataset MNIST selama 50 epochs menunjukkan proses konvergensi yang sangat cepat dan stabil dengan tingkat akurasi akhir mencapai 99%. Laporan klasifikasi membuktikan performa deteksi sempurna (recall 1.00) pada sebagian besar kelas numerik, mengonfirmasi bahwa model MoE sangat andal dalam mengenali pola garis dan sudut geometris tulisan tangan yang relatif sederhana.
3. Ketangguhan terhadap Kompleksitas Visual FashionMNIST: Pada dataset FashionMNIST yang memiliki variasi tekstur lebih rumit, model MoE mampu mempertahankan performa yang kuat dengan akurasi pengujian sebesar 90%. Meskipun terdapat kendala inheren berupa visual ambiguity (ambiguitas visual) beresolusi rendah pada kluster pakaian tubuh bagian atas—khususnya pada kelas Shirt (kemeja) yang memiliki tumpang tindih semantik dengan T-shirt, Pullover, dan Coat—model tetap meraih performa yang hampir sempurna (0.99) pada objek dengan siluet tegas seperti Trouser (celana) serta kelompok alas kaki.

5. SARAN

Untuk pengembangan penelitian lebih lanjut di masa mendatang, beberapa poin saran yang dapat direkomendasikan adalah:

1. Eksplorasi Skala dan Resolusi Citra: Penelitian selanjutnya disarankan untuk menguji arsitektur MoE heterogen ini pada dataset citra berwarna (RGB) dengan resolusi yang lebih tinggi dan skala besar (seperti CIFAR-100 atau ImageNet) untuk mengevaluasi batas skalabilitas Gating Network.
2. Optimalisasi Arsitektur Gating Network: Perlu dilakukan eksperimen lebih lanjut mengenai struktur jaringan gerbang, misalnya dengan menggunakan mekanisme Sparse Gating atau menambahkan komponen Top-K routing guna meningkatkan efisiensi waktu komputasi saat jumlah expert yang digunakan bertambah.

3. Analisis Kuantitatif Perilaku Routing: Disarankan untuk mengekstraksi dan memvisualisasikan log bobot gating secara statistik per kelas untuk memberikan grafik interpretabilitas yang lebih mendalam mengenai pola kapan model secara mutlak memilih fitur lokal CNN atau fitur global ViT.

DAFTAR PUSTAKA

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[5] A. G. Howard *et al*., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[6] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[7] S. Khan *et al*., "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–41, 2022.

[8] S. d'Ascoli *et al*., "ConViT: Improving vision transformers with soft convolutional inductive biases," in *Proc. International Conference on Machine Learning (ICML)*, 2021, pp. 2286–2296.

[9] A. Dosovitskiy *et al*., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[10] H. Touvron *et al*., "Training data-efficient image transformers & distillation through attention," in *Proc. International Conference on Machine Learning (ICML)*, 2021, pp. 10347–10357.

[11] Z. Liu *et al*., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[12] Z. Dai *et al*., "CoAtNet: Marrying convolution and attention for all data sizes," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 3965–3977, 2021.

[13] B. Graham *et al*., "LeViT: A vision transformer in ConvNet's clothing for faster inference," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12259–12269.

[14] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. International Conference on Learning Representations (ICLR)*, 2022.

[15] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[16] N. Shazeer *et al*., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[17] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 8583–8595, 2021.

[18] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal contrastive learning with LIMoE: The language-image mixture of experts," *arXiv preprint arXiv:2206.02770*, 2022.

[19] J. Puigcerver, C. Riquelme, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," in *Proc. International Conference on Learning Representations (ICLR)*, 2024.

[20] Y. Zhou *et al*., "Mixture-of-experts with expert choice routing," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 7103–7114, 2022.