BENCHMARKING VISION TRANSFORMER KLASIFIKASI VISUAL MASAKAN PADANG DENGAN ROBUSTNESS MELALUI AUGMENTASI DATA

BENCHMARKING VISION TRANSFORMER FOR VISUAL CLASSIFICATION OF PADANG CUISINE WITH ROBUSTNESS THROUGH DATA AUGMENTATION

Akmal Hisyam Pradhana[1], Erna Daniati[2*]
*E-mail : ernadaniati@unpkediri.ac.id

[1,2] Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri

Abstrak

Masakan Padang merupakan kuliner Indonesia dengan variasi visual yang kompleks. Penelitian ini mengembangkan sistem klasifikasi citra masakan Padang menggunakan Vision Transformer (ViT) dengan penguatan robustness melalui data augmentation. Dataset berjumlah 993 citra mencakup sembilan kelas populer. Lima varian ViT diuji, dan ViT-B/16 menghasilkan akurasi tertinggi 95%, diikuti ViT-L/16 (91%) dan ViT-H/14 (90%), sedangkan patch size besar menunjukkan akurasi lebih rendah. Augmentasi data terbukti meningkatkan generalisasi model, meski dataset terbatas. Evaluasi menunjukkan precision, recall, dan F1-score di atas 0.90 pada sebagian besar kelas. Hasil ini membuktikan ViT efektif dalam mengenali detail visual masakan Padang serta berpotensi untuk sistem klasifikasi makanan berbasis deep learning sekaligus mendukung pelestarian kuliner tradisional Indonesia.

Kata kunci**:** *Augmentasi Data, Klasifikasi Visual, Masakan Padang, Robustness, Vision Transformer.*

Abstract

Padang cuisine is an Indonesian culinary heritage known for its complex visual variations. This study develops a classification system for Padang cuisine images using the Vision Transformer (ViT) architecture, with robustness enhanced through data augmentation. The dataset consists of 993 images across nine popular classes. Five ViT variants were tested, with ViT-B/16 achieving the highest accuracy of 95%, followed by ViT-L/16 (91%) and ViT-H/14 (90%), while larger patch sizes yielded lower accuracy. Data augmentation proved effective in improving model generalization despite the limited dataset. Evaluation results show precision, recall, and F1-scores above 0.90 for most classes. These findings demonstrate that ViT is effective in recognizing the visual details of Padang cuisine and has strong potential for developing deep learning–based food classification systems, while also supporting the preservation of Indonesia's traditional culinary heritage.

Keywords*: Data Augmentation, Visual Classification, Padang Cuisine, Robustness, Vision Transformer.*

1. PENDAHULUAN

Salah satu makanan khas Indonesia yang telah mendunia dan memiliki cita rasa yang khas karena penggunaan rempah-rempahnya adalah masakan Padang. Hal ini dibuktikan oleh survei CNN International pada tahun 2011, di mana rendang, salah satu hidangan khas Padang, berhasil menempati peringkat pertama dalam daftar World's 50 Delicious Foods [1]. Masakan Padang adalah salah satu jenis masakan tradisional Indonesia yang terkenal di dunia, dikenal dengan kekayaan rasa pedas dan bumbu rempah yang kuat. Masakan ini yang berasal dari daerah Sumatera Barat, telah menjadi salah satu kuliner yang banyak diminati di berbagai penjuru dunia. Keunikan masakan Padang tidak hanya terletak pada rasa, tetapi juga pada cara penyajian dan variasi menu yang sangat beragam [2]. Perkembangan teknologi dan informasi pada era digital memberikan dampak signifikan yang merambah hampir semua bidang kehidupan manusia [3] [4]. Dalam era digital perkembangan teknologi kecerdasan buatan (AI) khususnya Deep learning dinilai mampu menyelesaikan identifikasi jenis masakan dalam bentuk gambar yang memiliki variasi visual tinggi. Dengan menggunakan deep learning, model dapat belajar pola-pola kompleks dari data gambar, membuat proses pengenalan masakan secara otomatis. Salah satu pendekatan yang digunakan dalam pengklasifikasian gambar menggunakan Deep Learning adalah Transfer Learning dengan Vision Transformer (ViT).

Meskipun banyak penelitian yang mengkaji pengenalan masakan menggunakan Teknik deep learning, belum banyak penelitian terdahulu yang secara spesifik fokus pada pengklasifikasian masakan Padang khususnya menggunakan Vision Transformer (ViT). Penelitian terdahulu menggunakan Vision Transformer (ViT), untuk klasifikasi kendaraan [5], [6], klasifikasi bidang medis [7], [8], [9], [10]. Klasifikasi flora & fauna [11], [12], [13], [14]. Dari penelitian terdahulu tersebut telah berhasil menggunakan metode Vision Transformer (ViT) dalam berbagai jenis tugas. Penelitian terkait klasifikasi citra makanan telah banyak dilakukan dalam beberapa tahun terakhir. Berbagai pendekatan berbasis Convolutional Neural Networks (CNN) telah berhasil menunjukkan performa yang baik dalam mengenali citra makanan secara umum [15], [16], [17]. Selain itu, pendekatan transfer learning juga banyak dimanfaatkan untuk meningkatkan akurasi pada dataset terbatas [18], [19]. Seiring berkembangnya arsitektur deep learning, Vision Transformer (ViT) mulai diperkenalkan sebagai alternatif CNN, yang mengadaptasi mekanisme self-attention dari NLP untuk tugas visi komputer [20]. Beberapa penelitian menunjukkan bahwa ViT dapat mencapai performa setara bahkan melebihi CNN [21], [22], [23], [24].

Sebagian besar penelitian klasifikasi makanan menggunakan VIT masih berfokus pada dataset internasional seperti Food-101, Food2K, CNFOOD-241 dengan kategori yang sudah banyak dieksplorasi [25], [26], [27], [28]. Belum banyak studi yang secara spesifik mengkaji klasifikasi makanan khas Indonesia, khususnya masakan Padang menggunakan Vision Transformer. Variasi tampilan visual dari masakan Padang misalnya tekstur gulai, warna rendang, hingga bentuk dendeng menjadi tantangan tersendiri yang belum banyak disentuh dalam literatur. Terdapat kesenjangan dalam penerapan arsitektur ViT untuk domain kuliner lokal Indonesia. Masih terbuka ruang penelitian yang secara spesifik fokus pada pengklasifikasian masakan Padang menggunakan Vision Transformer (ViT). Sehingga peneliti tertarik mengklasifikasi masakan Padang menggunakan Vision Transformer (ViT).

Sehingga tujuan dari penelitian untuk mengembangkan model klasifikasi gambar menggunakan Vision Transformer (ViT) dalam mengenali berbagai jenis masakan Padang Indonesia. Kontribusi penelitian yaitu melakukan klasifikasi citra gambar menggunakan dataset khusus masakan Padang yang mencakup beberapa kategori populer, menerapkan dan mengevaluasi performa Vision Transformer dalam mengenali citra masakan Padang, serta menganalisis tingkat akurasi

dan potensi pengembangan sistem klasifikasi makanan berbasis deep learning pada konteks kuliner indonesia. Nasi Padang terkenal di berbagai belahan dunia dan menjadi simbol dari kekayaan rasa dan keberagaman kuliner Indonesia. Dengan menggunakan teknologi pengenalan gambar untuk mengklasifikasikan berbagai jenis masakan Padang, Tidak hanya dapat memperkenalkan masakan tradisional Indonesia ke mancanegara, tetapi juga berkontribusi dalam upaya pelestarian warisan kuliner Indonesia yang memiliki nilai historis dan budaya yang tinggi dalam bidang Artificial intelligence.

2. METODOLOGI

2.2 Dataset

Memastikan bahwa data yang digunakan telah melalui pembersihan dan pengolahan sehingga menjadi data yang berkualitas, konsisten, dan siap digunakan dalam analisis lebih lanjut. Proses ini juga mencakup identifikasi data yang tidak relevan, penanganan data yang hilang, serta transformasi data agar sesuai dengan kebutuhan penelitian [29]. Dataset yang digunakan dalam proyek ini bersumber dari Kaggle dengan nama "Padang Cuisine (Indonesian Food Image Dataset)" Faldo Fajri Afrinanto (2022). Dataset gambar ini memiliki struktur terdiri dari 9 folder kelas yang di dalamnya terdapat file JPG seperti pada Gambar 3. Dataset dibagi menjadi dua set, pelatihan dan validasi dengan perbandingan 70:30 menggunakan fungsi train_test_split, sambil memastikan bahwa pembagian data ini dapat diulang dengan mengatur random_state=42.

Gambar 1. Dataset Gambar

9 folder kelas dalam data diantaranya yaitu Rendang Daging Sapi, Pop Ayam, Ayam Goreng, Dendeng Batokok, Kari Ikan, Kari Tambusu, Kari Tunjang, Telur Balado, dan Omelet Padang dengan total 993 gambar yang secara rinci seperti pada tabel 1.

Table 1. Daftar Class

Class	Total Dataset
Ayam Goreng	107
Ayam Pop	113
Daging Rendang	104
Dendeng Batokok	109
Gulai Ikan	111
Gulai Tambusu	103
Gulai Tunjang	119
Telur Balado	111
Telur Dadar	116

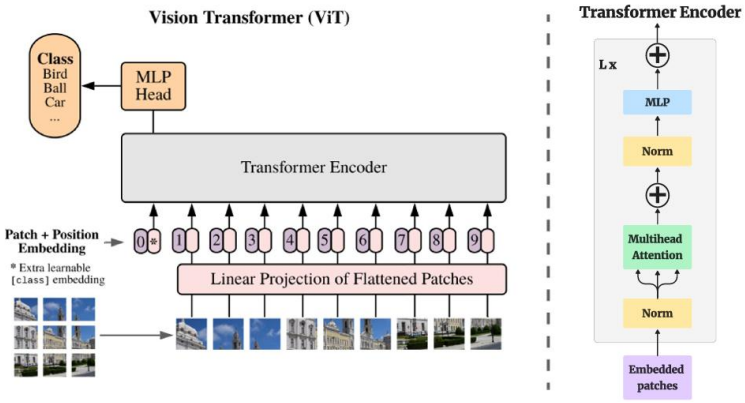
2.3 Data Augmentation

Data augmentation merupakan pendekatan kunci dalam deep learning untuk meningkatkan generalisasi model, khususnya Vision Transformer (ViT) yang memiliki bias induktif lemah dan by design lebih sensitif terhadap jumlah data pelatihan dan teknik augmentasi. Melalui modifikasi terkontrol seperti rotasi, flipping, cropping, dan jitter warna, teknik ini menambah variasi sintetik pada dataset, sehingga mencegah overfitting dan memperkuat ketahanan model terhadap data yang belum pernah dilihat sebelumnya [30] [31]. Pada tahap ini transform.Compose digunakan untuk menerapkan serangkaian transformasi pada dataset pelatihan dan validasi. Transformasi pada data pelatihan lebih kompleks untuk meningkatkan variasi melalui augmentasi, yang dapat membantu model untuk belajar dengan lebih baik dan menghindari overfitting.

Sementara itu, transformasi pada data validasi lebih sederhana agar evaluasi tetap konsisten dan representatif. Setelah itu, dataset pelatihan dan validasi dimuat menggunakan datasets. ImageFolder dengan transformasi yang telah diterapkan. Terakhir, DataLoader dibuat untuk dataset pelatihan dengan batch size 32, dan proses pengacakan data (shuffle) diterapkan pada train loader untuk memastikan model tidak terpengaruh oleh urutan data pelatihan, yang dapat menyebabkan bias. Tahapan ini sangat penting dalam pelatihan model deep learning, terutama untuk mengurangi bias model yang dapat muncul jika data tidak beragam. Dengan data yang lebih bervariasi, model menjadi lebih robust dan mampu menghadapi data baru yang belum pernah dilihat sebelumnya. Augmentasi juga meningkatkan efisiensi penggunaan data dan mengurangi ketergantungan pada proses pengumpulan data tambahan yang memakan biaya dan waktu. Pada akhirnya dataset yang lebih beragam, model yang lebih robust dalam mengenali pola pada data yang berbeda, serta performa model yang stabil pada data validasi dan pengujian. Transformasi sederhana yang diterapkan pada data validasi memastikan evaluasi yang konsisten dan representatif terhadap kualitas prediksi model, sehingga model dapat digunakan dengan kepercayaan lebih tinggi di aplikasi dunia nyata.

2.5 Model

Vision Transformer (ViT) merupakan arsitektur jaringan saraf yang mengadaptasi mekanisme Transformer yang awalnya dirancang untuk pemrosesan bahasa alami, ke dalam domain Deep Learning Computer Vision. Dalam implementasi arsitektur Vision Transformer (ViT) hanya memanfaatkan bagian encoder dari struktur Transformer. Encoder ini berfungsi untuk mengubah citra masukan menjadi representasi fitur yang lebih abstrak sehingga dapat digunakan dalam berbagai tugas visi komputer, khususnya klasifikasi gambar. Secara garis besar, mekanisme kerja ViT dimulai dengan memecah gambar menjadi potongan kecil berukuran sama (patch). Setiap patch kemudian diproyeksikan secara linear dan diberikan informasi posisi agar urutan spasial tetap terjaga. Selanjutnya, rangkaian vektor hasil embedding tersebut diproses melalui tumpukan encoder Transformer. Model ViT kemudian dilatih menggunakan label gambar, sehingga representasi yang diperoleh dapat dioptimalkan untuk menghasilkan prediksi pada tugas klasifikasi [32], [33].



Gambar 2. Arsitektur VIT Dalam Pengenalan Gambar

Dalam studi ini diterapkan lima varian arsitektur Vision Transformer (ViT), yaitu ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32, serta ViT-H/14 [21]. Perbandingan detail konfigurasi dari kelima model tersebut disajikan pada Tabel 1. Seluruh model ViT yang digunakan merupakan hasil pre-training pada dataset ImageNet-1K (yang terdiri dari sekitar satu juta gambar dengan 1.000 kelas) menggunakan implementasi dari PyTorchn dengan teknik SWAG (Stochastic Weight Averaging-Gaussian).

Table 2. Spesifikasi 5 Varian Arsitektur Vision Transformer (Vit)

Varian Model	Ukuran Patch (P)	Jumlah Layer (L)	Dimensi Hidden (D)	Ukuran MLP	Jumlah Head
ViT-B/16	16	12	768	3072	12
ViT-B/32	32	12	768	3072	12
ViT-L/16	16	24	1024	4096	16
ViT-L/32	32	24	1024	4096	16
ViT-H/14	14	32	1280	5120	16

Huruf B, L, dan H pada nama model masing-masing merepresentasikan kategori Base, Large, dan Huge. Istilah patch size merujuk pada ukuran potongan citra kecil yang digunakan sebagai unit masukan. Layers (L) mengindikasikan jumlah blok encoder yang terdapat pada Transformer. Sementara itu, hidden size (D) menggambarkan dimensi embedding yang digunakan untuk merepresentasikan fitur. MLP size menunjukkan jumlah unit tersembunyi pada lapisan MLP di dalam encoder Transformer. Adapun head merepresentasikan banyaknya kepala atensi yang digunakan pada mekanisme Multi-Head Self-Attention (MSA) dalam arsitektur Transformer.

2.6 Model Training

Konfigurasi hyperparameter yang digunakan secara konsisten pada seluruh eksperimen untuk lima varian Vision Transformer (ViT). Fungsi kerugian yang dipakai adalah Cross Entropy Loss, yang umum digunakan dalam permasalahan klasifikasi karena mampu mengukur perbedaan distribusi probabilistik antara prediksi model dan label sebenarnya. Optimisasi parameter dilakukan menggunakan algoritma Adam (Adaptive Moment Estimation), dengan tingkat pembelajaran (learning rate) sebesar 0.001, yang dianggap seimbang antara stabilitas konvergensi dan kecepatan pelatihan [34]. Ukuran batch ditetapkan sebesar 32, sehingga dalam setiap iterasi pelatihan, 32 sampel diproses secara paralel. Jumlah epochs ditentukan sebanyak 25 untuk memberikan waktu pelatihan yang memadai dalam memperoleh konvergensi model.

Table 3. Hyperpramter

Hyperparameter	Nilai / Keterangan
Loss Function	Cross Entropy Loss (nn.CrossEntropyLoss())
Optimizer	Adam (torch.optim.Adam)
Learning Rate	0.001
Batch Size	32
Epochs	25
Mode Training	model.train()
Mode Evaluasi	model.eval()
Gradien Reset	optimizer.zero_grad()
Forward Pass	outputs = model(inputs)
Backward Pass	loss.backward()
Update Bobot	optimizer.step()
No Grad Evaluasi	with torch.no_grad()
Metric Utama	Loss & Akurasi (Train & Validation)

Selama proses pelatihan mode yang digunakan adalah model.train(), sedangkan pada tahap evaluasi diterapkan model.eval() agar gradien tidak dihitung. Untuk menghindari akumulasi gradien yang tidak diinginkan, setiap iterasi dilakukan reset gradien (optimizer.zero_grad()), kemudian proses forward pass menghitung keluaran model, dilanjutkan dengan backward pass melalui propagasi balik menggunakan loss.backward(), dan pembaruan bobot dilakukan dengan optimizer.step(). Selama tahap evaluasi, perhitungan gradien dinonaktifkan menggunakan torch.no_grad() untuk meningkatkan efisiensi komputasi. Sebagai metrik utama, penelitian ini menggunakan kombinasi nilai kerugian (loss) dan akurasi pada data pelatihan maupun validasi. Dengan menjaga konsistensi hyperparameter pada kelima arsitektur ViT (ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32, dan ViT-H/14), perbandingan kinerja antar-model dapat dilakukan secara adil tanpa adanya pengaruh variasi dari konfigurasi pelatihan.

2.7 Evaluation

Proses penilaian kinerja model setelah model dilatih pada data pelatihan. Tujuan utama dari evaluasi ini adalah untuk mengukur seberapa baik model dalam memprediksi hasil yang benar dan untuk mengidentifikasi kekuatan serta kelemahan model ketika dihadapkan pada data yang belum pernah dilihat sebelumnya (data uji) [35]. Evaluasi dilakukan dengan menggunakan skenario yang telah direncanakan sebelumnya, guna menilai tingkat efektivitas & efisiensi sistem [36]. Salah satu cara untuk melakukan ini adalah dengan menghasilkan Classification Report, yang memberikan informasi lebih rinci mengenai kinerja model pada setiap kelas yang ada. Dalam classification report ini, kita dapat melihat nilai precision, recall, dan f1-score untuk setiap kelas [37]. Precision mengukur akurasi prediksi positif model, yaitu sejauh mana prediksi model yang dikategorikan positif benar-benar relevan. Recall, di sisi lain, mengukur sejauh mana model dapat menemukan semua instance positif yang ada dalam data. Sementara itu, F1-score adalah rata-rata harmonis antara precision dan recall, yang memberikan gambaran keseimbangan antara keduanya. Dengan menggunakan classification report, kita bisa menilai apakah model cenderung menghasilkan banyak false positives atau false negatives, serta seberapa baik model secara keseluruhan dalam mengklasifikasikan data. Selain itu Salah satu pendekatan yang dapat digunakan untuk menilai kinerja model adalah dengan menggunakan metode Confusion Matrix.

Metode ini mengukur kinerja klasifikasi dengan membandingkan hasil klasifikasi yang sebenarnya dengan hasil yang diprediksi oleh sistem. Ada empat istilah yang digunakan untuk mengukur kinerja klasifikasi, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Gambar 2 menunjukkan ilustrasi tabel confusion matrix [38].

		Predicted Class	
	Total Population n = a number	False (0)	True (1)
Actual Class	False (0)	TN True Negative	FP False Positive
	True (1)	FN False Negative	TP True Positive

Gambar 3. Table Confusion Matrix

3. HASIL DAN PEMBAHASAN

3.1 HASIL

3.1.1 Hasil 5 Model Vision Transformer

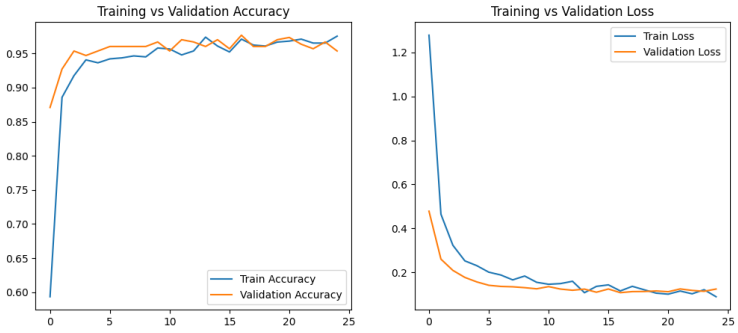
Penelitian menguji lima varian arsitektur Vision Transformer (ViT) yaitu ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32, dan ViT-H/14 yang seluruhnya telah melalui tahap pre-training pada ImageNet-1K dan kemudian dilakukan fine-tuning menggunakan dataset masakan Padang. Berdasarkan hasil evaluasi pada Tabel 4., ViT-B/16 berhasil memperoleh akurasi tertinggi sebesar 95%, diikuti oleh ViT-L/16 dengan akurasi 91%, dan ViT-H/14 dengan akurasi 90%. Sementara itu, varian ViT-B/32 dan ViT-L/32 menunjukkan performa lebih rendah dengan akurasi masing-masing 83% dan 81%.

Tabel 4. Evaluasi Classification Report

Model	Accuracy
ViT-B/16-in1k	**0.95**
ViT-B/32-in1k	0.83
ViT-L/16-in1k	0.91
ViT-L/32-in1k	0.81
ViT-H/14-in1k	0.90

3.1.2 Hasil Selama Training Model Terbaik

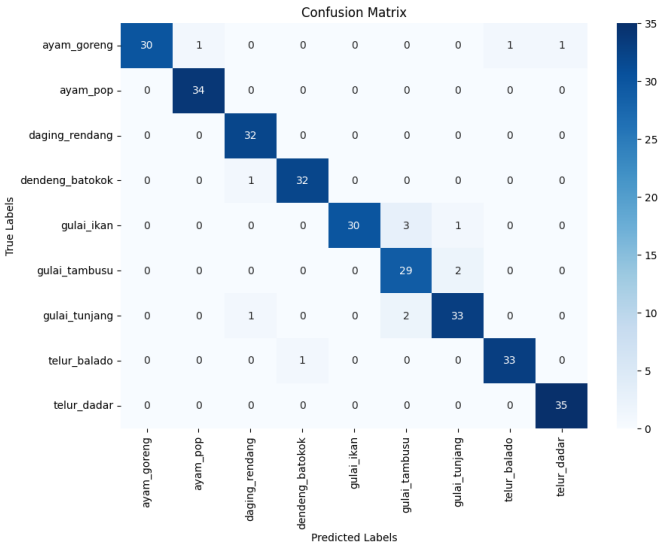
Hasil pelatihan model Vision Transformer (ViT) pada Gambar 4. merupakan model ViT-B/16 yang berhasil memperoleh akurasi tertinggi sebesar 95%. Selama 25 epoch pada dataset masakan Padang menunjukkan performa yang sangat baik. Grafik akurasi memperlihatkan bahwa pada epoch awal akurasi model masih berada di kisaran 60%, namun kemudian meningkat tajam hingga melampaui 90% pada epoch ke-5. Setelah itu, akurasi terus stabil hingga mencapai sekitar 95% pada akhir pelatihan. Pola yang sama juga terlihat pada akurasi validasi, yang hampir sejajar dengan akurasi pelatihan, menandakan bahwa model mampu melakukan generalisasi dengan baik terhadap data uji. Dari sisi loss, nilai awal yang cukup tinggi (>1.2) mengalami penurunan drastis dalam beberapa epoch pertama, kemudian stabil di bawah 0.2 setelah epoch ke-10. Perbedaan antara training loss dan validation loss relatif kecil sehingga tidak terlihat adanya gejala overfitting yang signifikan.



Gambar 4. Hasil visualisasi training dan validation

3.1.3 Hasil Confusion Matrix Model Terbaik

Hasil confusion matrix pada Gambar 5. yang diperoleh menunjukkan bahwa sebagian besar gambar berhasil diklasifikasikan dengan tepat. Namun, terdapat beberapa kesalahan klasifikasi, seperti ayam_goreng yang sesekali diprediksi sebagai ayam_pop, atau gulai_ikan yang beberapa kali diklasifikasikan sebagai gulai_tambusu dan gulai_tunjang. Kesalahan semacam ini kemungkinan disebabkan oleh kesamaan visual antara beberapa jenis makanan yang sulit dibedakan oleh model. Selain menggunakan confusion matrix, analisis lebih mendalam dapat dilakukan dengan membuat classification report seperti pada Tabel 5. untuk meninjau nilai precision, recall, dan f1-score setiap Class.



Gambar 5. Hasil confusion matrix

3.1.4 Hasil Classification Report Setiap Class

Evaluasi kinerja model Vision Transformer (ViT) pada klasifikasi citra masakan Padang menghasilkan classification report setiap class seperti pada Tabel 5. Secara umum, nilai precision, recall, dan F1-score menunjukkan performa yang sangat baik, dengan rata-rata di atas 0.90 pada hampir semua kelas. Beberapa kelas seperti ayam_pop dan telur_dadar mencapai nilai recall 1.00 dengan F1-score 0.99, yang menunjukkan bahwa model mampu mengenali hampir seluruh sampel dari kedua kelas tersebut tanpa kesalahan signifikan. Kelas ayam_goreng dan gulai_ikan memperoleh precision sempurna (1.00), meskipun recall gulai_ikan sedikit lebih

rendah (0.88), yang berarti masih ada sebagian kecil gambar gulai_ikan yang terklasifikasi ke kelas lain. Sebaliknya, performa terendah terlihat pada kelas gulai_tambusu, dengan precision 0.85 dan F1-score 0.89. Hal ini mengindikasikan bahwa model cukup sering salah mengklasifikasikan gambar gulai_tambusu ke kelas lain, kemungkinan disebabkan oleh kemiripan tekstur dan warna dengan jenis gulai lain seperti gulai_tunjang atau gulai_ikan. Secara keseluruhan, hasil evaluasi menunjukkan bahwa Vision Transformer (ViT) mampu mengenali variasi visual dari masakan Padang dengan tingkat akurasi yang sangat tinggi.

Tabel 5. Hasil Classification Report Setiap Class

Class	Precision	Recall	F1-Score
ayam_goreng	1.00	0.91	0.95
ayam_pop	0.97	1.00	0.99
daging_rendang	0.94	1.00	0.97
dendeng_batokok	0.97	0.97	0.97
gulai_ikan	1.00	0.88	0.94
gulai_tambusu	0.85	0.94	0.89
gulai_tunjang	0.92	0.92	0.92
telur_balado	0.97	0.97	0.97
telur_dadar	0.97	1.00	0.99

3.2 PEMBAHASAN

3.2.1 Pengaruh Perbedaan Model VIT & Data Augmentation

Berdasarkan hasil pada tabel 4 menunjukkan bahwa ukuran patch size memiliki pengaruh signifikan terhadap performa klasifikasi citra masakan Padang. Model dengan patch size yang lebih kecil (ViT-B/16 dan ViT-L/16) cenderung memberikan akurasi lebih tinggi dibandingkan patch size yang lebih besar (ViT-B/32 dan ViT-L/32). Hal ini dapat dijelaskan karena patch yang lebih kecil mampu menangkap detail visual yang lebih kaya, seperti tekstur rendang, warna kari, atau pola bumbu pada masakan Padang. Selain itu, meskipun ViT-H/14 memiliki kompleksitas arsitektur yang lebih besar dengan jumlah layer dan hidden dimension lebih tinggi, akurasi yang diperoleh (0.90) masih sedikit lebih rendah dibandingkan ViT-B/16 (0.95). Hal ini dapat dikaitkan dengan ukuran dataset yang relatif kecil (993 gambar). Model dengan kompleksitas sangat tinggi seperti ViT-H/14 memerlukan jumlah data yang lebih besar agar dapat mencapai performa optimal.

Penggunaan data augmentation terbukti berperan penting dalam meningkatkan generalisasi model. Teknik augmentasi seperti rotasi, flipping, dan jitter warna membuat model lebih robust terhadap variasi visual dalam masakan Padang, sehingga dapat mengurangi risiko overfitting. Dari segi arsitektur, ViT-B/16 menunjukkan trade-off yang seimbang antara kompleksitas model dan kapasitas dataset, sehingga mampu menghasilkan performa terbaik. Sementara itu, performa ViT-L/32 dan ViT-B/32 yang lebih rendah menegaskan bahwa patch size yang terlalu besar justru membuat model kehilangan detail penting dalam citra, yang sangat krusial untuk membedakan antar kelas makanan Padang yang memiliki kesamaan visual tinggi. Secara keseluruhan, hasil ini memperlihatkan bahwa Vision Transformer (ViT) memiliki potensi yang kuat untuk digunakan dalam klasifikasi makanan lokal Indonesia, khususnya masakan Padang. Dengan capaian akurasi hingga 95%, penelitian ini membuktikan bahwa pendekatan berbasis ViT mampu mengenali karakteristik unik masakan tradisional Indonesia dengan baik.

3.2.2 Pengaruh Transfer Learning Pada Model VIT

Hasil performa selama pelatihan membuktikan bahwa arsitektur Vision Transformer (ViT) dengan konfigurasi hyperparameter yang digunakan, yaitu Cross Entropy Loss, optimizer Adam, learning rate 0.001, batch size 32, dan 25 epoch, mampu memberikan performa klasifikasi yang optimal. Keberhasilan ini tidak lepas dari penerapan Transfer Learning, di mana model telah melalui pre-training pada ImageNet-1K sebelum dilakukan fine-tuning pada dataset masakan Padang. Pendekatan tersebut terbukti mempercepat proses konvergensi model, terlihat dari peningkatan akurasi yang signifikan pada epoch-epoch awal. Selain itu, penggunaan teknik data augmentation seperti rotasi, flipping, dan jitter warna juga berperan penting dalam meningkatkan keragaman data, sehingga model tidak sekadar menghafal pola visual tertentu melainkan mampu belajar fitur yang lebih umum.

Secara keseluruhan hasil pelatihan memperlihatkan bahwa model mencapai stabilitas setelah epoch ke-10 dengan akurasi dan loss yang konsisten baik pada data pelatihan maupun validasi. Kondisi ini menunjukkan bahwa jumlah epoch 25 sudah cukup untuk menghasilkan performa optimal. Tidak adanya perbedaan signifikan antara train accuracy dan validation accuracy, serta kesamaan pola penurunan train loss dan validation loss, semakin menegaskan bahwa model memiliki kemampuan generalisasi yang baik meskipun dataset yang digunakan relatif terbatas jumlahnya (993 gambar). Dengan capaian tersebut, dapat disimpulkan bahwa Vision Transformer (ViT) merupakan pendekatan yang sangat potensial untuk klasifikasi citra makanan khas Indonesia, khususnya masakan Padang, karena tidak hanya menghasilkan akurasi tinggi tetapi juga konsistensi performa pada data uji.

4. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian dapat disimpulkan bahwa Vision Transformer (ViT) memiliki potensi besar dalam klasifikasi citra masakan Padang yang memiliki variasi visual kompleks. Dari lima varian arsitektur yang diuji, ViT-B/16 menunjukkan performa terbaik dengan akurasi mencapai 95%, diikuti oleh ViT-L/16 dan ViT-H/14 yang masing-masing memperoleh akurasi 91% dan 90%. Hasil ini menegaskan bahwa ukuran patch yang lebih kecil memberikan keunggulan dalam menangkap detail tekstur dan pola visual khas masakan Padang dibandingkan patch yang lebih besar. Selain itu, penggunaan teknik data augmentation terbukti efektif dalam meningkatkan generalisasi model dan mencegah overfitting meskipun jumlah dataset relatif terbatas. Proses pelatihan menunjukkan bahwa model mampu mencapai stabilitas dengan akurasi tinggi baik pada data latih maupun validasi, menandakan kemampuan generalisasi yang baik. Dengan demikian, penelitian ini tidak hanya berhasil membuktikan efektivitas ViT dalam pengenalan masakan Padang, tetapi juga membuka peluang pengembangan sistem klasifikasi makanan berbasis deep learning untuk mendukung pelestarian dan promosi kuliner tradisional Indonesia di era digital.

5. DAFTAR RUJUKAN

[1] M. D. Hadi Pratama, K. Al Kautsar, R. Hidayat, V. Melistiana, dan Tiarapuspa, “ANALISIS BISNIS STRATEGI NASI PADANG 99,” *Jurnal Ekonomi Trisakti*, vol. 3, no. 1, hlm. 601–610, Jan 2023, doi: 10.25105/jet.v3i1.15519.

[2] Lulu Kamelia, Melanda Wulandari, Santi Pertiwi Hari Sandi, dan Dwi Epty Hidayaty, “Peran Kinerja Karyawan Pada Peningkatan Produktivitas Rumah Makan Padang Alam Minang,” *Journal of Management and Creative Business*, vol. 1, no. 3, hlm. 190–198, Jun 2023, doi: 10.30640/jmcbus.v1i3.1180.

[3] Ferry Kurniawan, Erlin Ayu Khrisnawati, Rizka Hadiwiyanti, dan Anindo Saka Fitri, "PENGUJIAN SISTEM INFORMASI MANAJEMEN SISWA BERBASIS WEBSITE MENGGUNAKAN METODE BLACK BOX DAN WHITE BOX," *Prosiding Seminar Nasional Teknologi dan Sistem Informasi*, vol. 2, no. 1, hlm. 249–261, Sep 2022, doi: 10.33005/sitasi.v2i1.306.

[4] K. A. Zahra Salsabilla, Tasya Diva Fortuna Hadi, Widya Pratiwi, dan Siti Mukaromah, "PENGARUH PENGGUNAAN KECERDASAN BUATAN TERHADAP MAHASISWA DI PERGURUAN TINGGI," *Prosiding Seminar Nasional Teknologi dan Sistem Informasi*, vol. 3, no. 1, hlm. 168–175, Nov 2023, doi: 10.33005/sitasi.v3i1.371.

[5] Y. Taki dan E. Zemmouri, "Vehicle Image Classification Method Using Vision Transformer," 2023, hlm. 221–230. doi: 10.1007/978-3-031-43520-1_19.

[6] X. Dong, P. Shi, Y. Tang, L. Yang, A. Yang, dan T. Liang, "Vehicle Classification Algorithm Based on Improved Vision Transformer," *World Electric Vehicle Journal*, vol. 15, no. 8, hlm. 344, Jul 2024, doi: 10.3390/wevj15080344.

[7] A. Halder, S. Gharami, P. Sadhu, P. K. Singh, M. Woźniak, dan M. F. Ijaz, "Implementing vision transformer for classifying 2D biomedical images," *Sci Rep*, vol. 14, no. 1, hlm. 12567, Mei 2024, doi: 10.1038/s41598-024-63094-9.

[8] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, dan A. Ayatollahi, "MedViT: A robust vision transformer for generalized medical image classification," *Comput Biol Med*, vol. 157, hlm. 106791, Mei 2023, doi: 10.1016/j.compbiomed.2023.106791.

[9] V. Nikitin dan N. Shapoval, "Vision transformer for skin cancer classification," *InterConf*, no. 33(155), hlm. 449–460, Mei 2023, doi: 10.51582/interconf.19-20.05.2023.039.

[10] O. A. Supriadi, E. Utami, dan D. Ariatmanto, "Deteksi Tumor Otak Melalui Gambar MRI Berdasarkan Vision Transformers dengan Tensorflow dan Keras," *Jurnal Informatika Universitas Pamulang*, vol. 8, no. 3, hlm. 385–392, Sep 2023, doi: 10.32493/informatika.v8i3.32707.

[11] C. Yang *dkk.*, "FishAI : Automated hierarchical marine fish image classification with vision transformer," *Engineering Reports*, vol. 6, no. 12, Des 2024, doi: 10.1002/eng2.12992.

[12] B. Gong, K. Dai, J. Shao, L. Jing, dan Y. Chen, "Fish-TViT: A novel fish species classification method in multi water areas based on transfer learning and vision transformer," *Heliyon*, vol. 9, no. 6, hlm. e16761, Jun 2023, doi: 10.1016/j.heliyon.2023.e16761.

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

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

[15] L. Bu, C. Hu, dan X. Zhang, "Recognition of food images based on transfer learning and ensemble learning," *PLoS One*, vol. 19, no. 1, hlm. e0296789, Jan 2024, doi: 10.1371/journal.pone.0296789.

[16] D. Liu, E. Zuo, D. Wang, L. He, L. Dong, dan X. Lu, "Deep Learning in Food Image Recognition: A Comprehensive Review," *Applied Sciences*, vol. 15, no. 14, hlm. 7626, Jul 2025, doi: 10.3390/app15147626.

[17] G. A. Tahir dan C. K. Loo, "A Comprehensive Survey of Image-Based Food Recognition and Volume Estimation Methods for Dietary Assessment," *Healthcare*, vol. 9, no. 12, hlm. 1676, Des 2021, doi: 10.3390/healthcare9121676.

[18] Y. Park, A.-C. Hauschild, dan D. Heider, "Transfer learning compensates limited data, batch effects and technological heterogeneity in single-cell sequencing," *NAR Genom Bioinform*, vol. 3, no. 4, Okt 2021, doi: 10.1093/nargab/lqab104.

[19] L. Alzubaidi *dkk.*, "Novel Transfer Learning Approach for Medical Imaging with Limited Labeled Data," *Cancers (Basel)*, vol. 13, no. 7, hlm. 1590, Mar 2021, doi: 10.3390/cancers13071590.

[20] K. Han *dkk.*, "A Survey on Vision Transformer," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 1, hlm. 87–110, Jan 2023, doi: 10.1109/TPAMI.2022.3152247.

[21] A. Dosovitskiy *dkk.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun 2021.

[22] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, dan A. Dosovitskiy, "Do Vision Transformers See Like Convolutional Neural Networks?," Mar 2022.

[23] J. H. L. Goh *dkk.*, "Comparative Analysis of Vision Transformers and Conventional Convolutional Neural Networks in Detecting Referable Diabetic Retinopathy," *Ophthalmology Science*, vol. 4, no. 6, hlm. 100552, Nov 2024, doi: 10.1016/j.xops.2024.100552.

[24] J. Maurício, I. Domingues, dan J. Bernardino, "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review," *Applied Sciences*, vol. 13, no. 9, hlm. 5521, Apr 2023, doi: 10.3390/app13095521.

[25] X. Gao, Z. Xiao, dan Z. Deng, "High accuracy food image classification via vision transformer with data augmentation and feature augmentation," *J Food Eng*, vol. 365, hlm. 111833, Mar 2024, doi: 10.1016/j.jfoodeng.2023.111833.

[26] T. Ghosh dan E. Sazonov, "Improving Food Image Recognition with Noisy Vision Transformer," Mar 2025.

[27] B. N. Jagadesh *dkk.*, "Enhancing food recognition accuracy using hybrid transformer models and image preprocessing techniques," *Sci Rep*, vol. 15, no. 1, hlm. 5591, Feb 2025, doi: 10.1038/s41598-025-90244-4.

[28] W. Min *dkk.*, "Large Scale Visual Food Recognition," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 8, hlm. 9932–9949, Agu 2023, doi: 10.1109/TPAMI.2023.3237871.

[29] S. Kurniawan, W. Gata, D. A. Puspitawati, N. -, M. Tabrani, dan K. Novel, "Perbandingan Metode Klasifikasi Analisis Sentimen Tokoh Politik Pada Komentar Media Berita Online," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 2, hlm. 176–183, Agu 2019, doi: 10.29207/resti.v3i2.935.

[30] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, dan L. Beyer, "How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers," Jun 2022.

[31] R. Bintang dan Y. Azhar, "Implementasi Data Augmentation untuk Klasifikasi Sampah Organik dan Non Organik Menggunakan Inception-V3," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 9, no. 3, hlm. 192–204, Sep 2024, doi: 10.14421/jiska.2024.9.3.192-204.

[32] K. Al-hammuri, F. Gebali, A. Kanan, dan I. T. Chelvan, "Vision transformer architecture and applications in digital health: a tutorial and survey," *Vis Comput Ind Biomed Art*, vol. 6, no. 1, hlm. 14, Jul 2023, doi: 10.1186/s42492-023-00140-9.

[33] A. M. Ali, B. Benjdira, A. Koubaa, W. El-Shafai, Z. Khan, dan W. Boulila, "Vision Transformers in Image Restoration: A Survey," *Sensors*, vol. 23, no. 5, hlm. 2385, Feb 2023, doi: 10.3390/s23052385.

[34] I. Kandel, M. Castelli, dan A. Popovič, "Comparative Study of First Order Optimizers for Image Classification Using Convolutional Neural Networks on Histopathology Images," *J Imaging*, vol. 6, no. 9, hlm. 92, Sep 2020, doi: 10.3390/jimaging6090092.

[35] T. A. Yoga Siswa dan Naufal Azmi Verdikha, "KOMPARASI ALGORITMA KLASIFIKASI UNTUK MENENTUKAN EVALUASI KINERJA TERBAIK PADA STATUS AKREDITASI SEKOLAH/MADRASAH KALIMANTAN TIMUR BERDASARKAN IASP 2020," *Jurnal Informatika, Teknologi dan Sains*, vol. 4, no. 3, hlm. 185–192, Agu 2022, doi: 10.51401/jinteks.v4i3.1807.

[36] E. Daniati, S. Sucipto, A. S. Wardani, dan A. H. Pradhana, "Usability Test on the System Determination Decision Support ReleaseProduct Towards Contribution Level Decision Maker," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 24, no. 3, hlm. 519–530, Jul 2025, doi: 10.30812/matrik.v24i3.3789.

[37] E. Daniati dan H. Utama, "ANALISIS SENTIMEN DENGAN PENDEKATAN ENSEMBLE LEARNING DAN WORD EMBEDDING PADA TWITTER," *Journal of Information System Management (JOISM)*, vol. 4, no. 2, hlm. 125–131, Jan 2023, doi: 10.24076/joism.2023v4i2.973.

[38] F. M. Fathoni, C. A. Putra, dan A. L. Nurlaili, "KLASIFIKASI PENYAKIT DAUN ANGGUR MENGGUNAKAN METODE K-NEAREST NEIGHBOR BERDASARKAN GRAY LEVEL CO-OCCURRENCE MATRIX," *Biner : Jurnal Ilmiah Informatika dan Komputer*, vol. 3, no. 1, hlm. 8–15, Jan 2024, doi: 10.32699/biner.v3i1.6332.