Klasifikasi Makanan Tradisional Indonesia dengan Semi Supervised Learning Menggunakan CNN dan Transformer

Ahmad Naufal Farras[1], Jeremy Mattathias Mboe[2], Kayla Riza Putri Irjayanto[3]

[1]Fakultas Teknologi Elektro dan Informatika Cerdas Institut Teknologi Sepuluh Nopember, Surabaya, email: 5054241038@student.its.ac.id
[2]Fakultas Teknologi Elektro dan Informatika Cerdas Institut Teknologi Sepuluh Nopember, Surabaya, email: 5054241012@student.its.ac.id
[3]Fakultas Teknologi Elektro dan Informatika Cerdas Institut Teknologi Sepuluh Nopember, Surabaya, email: 5054241028@student.its.ac.id
Corresponding Author: Ahmad Naufal Farras

ABSTRAK — Penggunaan *machine learning* sudah sangat umum digunakan oleh sebagian besar industri di era sekarang. Salah satu metode machine learning yang kerap digunakan adalah *semi supervised learning* untuk melabelkan data secara manual lalu memprediksi dengan model yang telah ditentukan sebelumnya. Pada studi kasus *label discovery* pada kumpulan gambar makanan tradisional Indonesia teknik *pseudo-labeling* digunakan pada dataset makanan tradisional Indonesia dan dikombinasikan dengan arsitektur *Convolutional Neural Network* (CNN) dan *Vision Transformer* (ViT) dengan model *pretrained* yang berbeda, seperti ResNet untuk CNN, dan DINO untuk ViT. Penelitian ini bertujuan untuk mengklasifikasikan kumpulan makanan tradisional Indonesia dengan akurasi tinggi, terutama dengan mengevaluasi kinerja dan efektivitas *pseudo-labeling* dalam mengidentifikasi makanan tradisional Indonesia pada data baru, serta menentukan arsitektur *pretrained* yang paling optimal di antara model yang telah kami uji. Metodologi yang diterapkan meliputi *pseudo-labeling* pada sebagian besar citra per kelas makanan tradisional Indonesia, di mana *confidence threshold* yang ketat ditetapkan untuk memastikan kualitas label semu. Proses ini dilanjutkan dengan penggunaan *pretrained* model untuk melakukan *fine-tuning* secara iteratif, yang kemudian digunakan untuk memprediksi data baru. Kami mengevaluasi kinerja seluruh model dengan metrik akurasi pada *test set* yang berlabel secara manual. Hasil penelitian menunjukkan bahwa implementasi *pseudo-labeling* berhasil meningkatkan pemanfaatan data tanpa label secara signifikan dalam klasifikasi. Studi ini menyimpulkan bahwa efektivitas *pseudo-labeling* sangat dipengaruhi oleh kualitas dan jenis arsitektur *pretrained*, memberikan kontribusi signifikan terhadap pengembangan sistem klasifikasi kuliner nasional yang kuat. Berdasarkan solusi usulan dan metode yang diberikan model terbaik yang kami dapatkan adalah DINO v2 dengan training accuracy 100.00 %, validation accuracy 93.45 % dan test accuracy 93.67%

KATA KUNCI — Klasifikasi Citra Makanan Tradisional Indonesia, Linear Probing, DINOv2, Time-Test Augmentation, Vision Transformer, Layer-wise Learning Rate Decay

I. PENDAHULUAN

Metode Machine Learning (ML) menjadi teknologi utama untuk inovasi industri di zaman sekarang. Namun, praktiknya sering terhambat oleh keterbatasan biaya dan waktu tinggi serta data yang sudah berlabel khususnya pada kasus klasifikasi citra. Tantangan ini muncul karena kemampuan model awal untuk melakukan *clustering* citra yang optimal masih terbatas, sehingga mengharuskan pelabelan data secara manual yang tidak efisien, terutama untuk *dataset* berskala besar. Keterbatasan data berlabel ini sangat terasa pada domain spesifik makanan tradisional Indonesia, di mana *dataset* yang tersedia masih minim.

Untuk mengatasi permasalahan ini, penelitian ini mengadopsi pendekatan *Semi Supervised Learning* dengan melakukan pelabelan data secara manual. Penelitian ini bertujuan untuk melakukan evaluasi komparatif kinerja model-model *pretrained deep learning* untuk klasifikasi gambar makanan tradisional Indonesia, seperti ResNet[1], Vision Transformer (ViT)[2], dan DINO[3], untuk menentukan konfigurasi optimal dalam tugas *label discovery* citra makanan tradisional Indonesia.

II. STUDI LITERATUR

Pengembangan model klasifikasi makanan secara otomatis memerlukan pemahaman yang mendalam terhadap konsep konsep dasar dari dari solusi yang diusulkan. Kajian ini mencakup teori-teori utama seperti Image classification menggunakan ViT B, ViT L, DINO v2.

A. KONSEP DATA MINING

Data mining merupakan salah satu metode machine learning yang komprehensif untuk mencari, mengumpulkan, menyaring, dan menganalisis data guna mengekstraksi pengetahuan yang berguna dari sebuah dataset yang bertujuan untuk menemukan relevansi, korelasi, atau informasi yang sebelumnya tidak diketahui dan tersembunyi di dalam dataset [4]. Dalam klasifikasi makanan tradisional Indonesia, penerapan data mining adalah dengan memanfaatkan ekstraksi fitur, penemuan pola, dan otomatisasi klasifikasi guna mengelompokkan dan membedakan jenis-jenis makanan berdasarkan karakteristik data citra yang ada.

B. IMAGE CLASSIFICATION DENGAN MODEL CNN DAN TRANSFORMER

Dalam klasifikasi citra, dua arsitektur utama digunakan: Convolutional Neural Network (CNN) dan Vision Transformer (ViT). CNN beroperasi berdasarkan prinsip konvolusi untuk mengekstrak fitur lokal dan hirarkis, memanfaatkan *inductive bias* bawaan seperti *translation equivariance*. Kelebihannya terletak pada efisiensi komputasi untuk menangkap struktur lokal dan objek kecil. Sementara itu, Transformer (seperti ViT) bekerja menggunakan mekanisme self-attention untuk memproses citra sebagai urutan *patches* (token). ViT unggul dalam menangkap ketergantungan jarak jauh (*long-range dependencies*) dan menunjukkan skalabilitas luar biasa pada *dataset* yang sangat besar, menjadikannya *backbone* yang sangat *robust* untuk fitur global. Tujuan dari kedua *backbone* ini pada dasarnya sama: menyediakan

representasi fitur yang kemudian diteruskan ke lapisan klasifikasi akhir, biasanya berupa lapisan *Fully Connected* yang menghasilkan probabilitas kelas melalui fungsi Softmax.

C. LAYER WISE LEARNING RATE DECAY

Layer-wise Learning Rate Decay (LLRD) adalah sebuah strategi pengoptimalan yang memberikan *learning rate* yang lebih kecil untuk lapiran Transformer yang lebih dekat ke input dan *learning rate* yang lebih besar untuk lapisan yang lebih dekat ke output. Teknik training ini bertujuan untuk menjaga stabilitas pelatihan pada *layer* yang sangat dalam dan memanfaatkan pengetahuan yang telah dipelajari dari *pre-training*. Dengan menggunakan *learning rate* yang tinggi pada *weight layer* awal, LLRD memastikan konvergensi yang lebih stabil dan efektif, sebuah praktik yang dipelopori dalam BERT [5], lalu diadopsi dalam Vision Transformers (ViT) [2].

D. TEST TIME AUGMENTATION

Test-Time Augmentation (TTA) adalah teknik pengujian yang efektif di mana satu gambar input dapat diubah menjadi beberapa versi yang telah diaugmentasi (misalnya dipotong, diputar, dan secara acak mengubah properti warna gambar) sebelum dimasukkan ke dalam model. Alasan utama memilih TTA adalah untuk meningkatkan keandalan dan akurasi prediksi model dengan mengevaluasi *input* dari berbagai perspektif. TTA bertujuan untuk mendapatkan hasil akhir yang lebih *robust* dengan mengambil rata-rata ataupun menggabungkan hasil prediksi dari semua versi yang diaugmentasi tersebut yang secara signifikan mengurangi kesalahan prediksi tanpa perlu melatih ulang model [6].

III. SOLUSI DAN USULAN

Untuk mengatasi tingginya biaya pelabelan data dan keterbatasan *dataset* yang masif, penelitian ini mengadopsi model DINOv2 berbasis Vision Transformer (ViT) yang dilatih melalui *Self-Supervised Learning* (SSL), yang berfungsi sebagai *feature extractor* yang unggul. DINO v2 menghasilkan fitur CLS *token* yang *robust*, yang secara efisien mendukung pendekatan Semi-Supervised Learning dan secara signifikan meminimalkan ketergantungan pada *dataset* berlabel dalam jumlah besar. Implementasi model dilakukan melalui *pipeline* pelatihan dua tahap, yaitu Linear Probing dan Full Fine-Tuning. Stabilitas *backbone* ViT yang sangat dalam ditingkatkan dengan teknik Layer-wise Learning Rate Decay (LLRD). Sementara itu, efisiensi dan keakuratan proses dijamin melalui Hyperparameter Tuning sistematis menggunakan Optuna untuk mencari konfigurasi parameter terbaik (*learning rate, weight decay*) di setiap tahap.

Selanjutnya, untuk menjamin kinerja optimal dan meningkatkan akurasi dalam penelitian ini, diusulkan penerapan Test-Time Augmentation (TTA) pada tahap evaluasi akhir. TTA berfungsi untuk secara signifikan meningkatkan keandalan dan akurasi prediksi model dengan merata-ratakan hasil dari beberapa versi *input* yang diaugmentasi. Hasil penelitian ini tidak hanya memberikan konfigurasi model *deep learning* (ViT) yang efisien dan berkinerja tinggi untuk klasifikasi citra makanan tradisional Indonesia, tetapi juga menghasilkan dasar teknologi yang kuat.

IV. METODE PENELITIAN

A. DATASET

Dataset yang digunakan berisi citra digital dari 15 jenis makanan tradisional Indonesia yang diperoleh melalui proses scraping menggunakan Google Images. Pada folder train, data tidak memiliki label, sehingga peserta diminta untuk melakukan pelabelan (labeling) atau sitasi secara mandiri guna membangun dataset berlabel. Dataset berlabel tersebut nantinya akan digunakan untuk melatih model yang kemudian dapat memprediksi label pada data di folder test.

1) DATASET TRAINING

Dataset yang berisi kumpulan gambar makanan tradisional Indonesia tanpa label, yang perlu dilakukan proses pelabelan atau sitasi sebelum digunakan untuk pelatihan model

Gambar 1. Contoh Data citra pada dataset train

2) DATASET TESTING

Dataset yang berisi kumpulan gambar makanan tradisional Indonesia tanpa label, yang akan digunakan untuk pengujian (prediksi) menggunakan model yang telah dilatih.

Gambar 2. Contoh data citra pada dataset test

B. PELATIHAN MODEL DINO v2

Training model dilakukan menggunakan backbone DINOv2-Large, yang merupakan model Vision Transformer (ViT) berbasis self-supervised learning. Model ini dilatih menggunakan framework PyTorch dan Hugging Face Transformers. Proses pelatihan dilakukan dalam dua tahap, yaitu Linear Probing dan Full Fine-Tuning, untuk memaksimalkan performa transfer learning. Pada tahap preprocessing, gambar diubah ukurannya menjadi 224 x 224 pixel dan dinormalisasi menggunakan mean dan standar deviasi dari dataset ImageNet. Konfigurasi untuk training DINOv2 adalah sebagai berikut:

- Model: DINOv2-Large (ViT-L/16)
- Tahap Pelatihan: Linear Probing (5 epoch) dan Full Fine-Tuning (45 epoch)
- Batch Size: 8
- Optimizer: AdamW
- Learning Rate:

- Linear Probing: 1e-3
- Fine-Tuning: 1e-4 (head), 5e-6 (backbone)
- Loss Function: CrossEntropyLoss dengan Label Smoothing (0.1)
- Regularisasi: Layer-wise Learning Rate Decay (LLRD) dengan decay rate 0.9
- Metode Validasi: Accuracy dan Loss

C. HYPERPARAMETER TUNING DENGAN OPTUNA

Proses hyperparameter tuning dilakukan menggunakan framework Optuna untuk mencari kombinasi hyperparameter terbaik yang dapat meningkatkan performa model. Tuning dilakukan untuk kedua tahap pelatihan, yaitu Linear Probing dan Full Fine-Tuning. Proses ini menggunakan pendekatan Bayesian Optimization dengan Median Pruner untuk menghentikan trial yang tidak menjanjikan lebih awal. Konfigurasi untuk hyperparameter tuning adalah sebagai berikut dengan jumlah trial 15 trial per tahap dan Hyperparameter yang Dicari:

Linear Probing:
- Learning Rate: [1e-4, 5e-3] (log scale)
- Weight Decay: [0.001, 0.1] (log scale)
- Batch Size: [4, 8, 12, 16]

Fine-Tuning:
- Head Learning Rate: [5e-5, 5e-4] (log scale)
- Backbone Learning Rate: [1e-6, 1e-5] (log scale)
- LLRD Decay Rate: [0.85, 0.95]
- Weight Decay: [0.001, 0.05] (log scale)
- Label Smoothing: [0.0, 0.2]
- Batch Size: [4, 8, 12]
- Metode Validasi: Accuracy dan Loss

D. ALUR KERJA SISTEM KLASIFIKASI MAKANAN

Alur Kerja sistem ini dirancang sebagai sebuah pipeline pemrosesan yang terdiri dari beberapa langkah sekuensial seperti pada flowchart di Gambar 3. Sistem menerima input berupa data citra makanan nusantara dan menguluarkan output klasifikasi dari nama makanan.

1) Input Gambar Makanan
Gambar makanan khas Indonesia diunggah dalam format RGB. Dataset terdiri dari 15 kelas makanan seperti ayam bakar, rendang, dan nasi goreng.

2) Preprocessing dan Augmentasi Data
Gambar makanan diproses menggunakan transformasi seperti:
RandomResizedCrop: Memotong gambar secara acak dengan skala 80%-100%.
RandomHorizontalFlip: Membalik gambar secara horizontal.
RandomRotation: Memutar gambar hingga ±15 derajat.
ColorJitter: Mengubah kecerahan, kontras, saturasi, dan hue.
Normalize: Menormalkan gambar menggunakan statistik ImageNet.

3) Ekstraksi Fitur dengan DINOv2
Gambar makanan diproses melalui backbone DINOv2-Large untuk menghasilkan representasi fitur global menggunakan token CLS (class token). Representasi ini digunakan sebagai input untuk lapisan klasifikasi.

4) Linear Probing (Tahap 1)
Pada tahap ini, backbone DINOv2 dibekukan (freeze) sehingga hanya lapisan klasifikasi (classifier head) yang dilatih. Tujuannya adalah untuk menyesuaikan classifier head dengan dataset makanan Indonesia.

5) Fine-Tuning (Tahap 2)
Setelah tahap Linear Probing selesai, seluruh model (backbone + head) dilatih menggunakan teknik Layer-wise Learning Rate Decay (LLRD). Teknik ini memastikan layer yang lebih dalam memiliki learning rate lebih kecil, sehingga penyesuaian backbone berjalan aman.

6) Hyperparameter Tuning dengan Optuna
Hyperparameter seperti learning rate, weight decay, dan batch size dioptimalkan menggunakan Optuna. Proses ini dilakukan untuk kedua tahap pelatihan (Linear Probing dan Fine-Tuning) guna menemukan kombinasi parameter terbaik.

7) Inferensi dan Prediksi Kelas
Model memprediksi kelas makanan untuk setiap gambar, menghasilkan probabilitas untuk setiap kelas. Jika Test-Time Augmentation (TTA) diaktifkan, prediksi dilakukan dengan berbagai augmentasi, dan hasilnya dirata-rata.

8) Evaluasi Model
Hasil prediksi dibandingkan dengan label ground truth menggunakan metrik seperti accuracy, precision, recall, dan confusion matrix. Analisis kesalahan dilakukan untuk mengidentifikasi kelas yang sering tertukar.

9) Output Akhir
Sistem menampilkan hasil klasifikasi berupa label makanan dan tingkat kepercayaan (confidence score) untuk setiap gambar. Hasil evaluasi seperti akurasi dan confusion matrix juga disimpan untuk analisis lebih lanjut.

V. HASIL DAN ANALISIS PENGUJIAN

A. HASIL EKSPERIMEN

Semua kode program pengembangan dibuat dalam bahasa pemrograman Python dengan menggunakan *framework* PyTorch untuk memproses data, mengaugmentasi data, melatih model, dan mengevaluasi model. Pengembangan model dilakukan dilakukan di tiga lingkungan komputasi yang berbeda, yakni Google Colaboratory, Kaggle Notebooks, dan komputer lokal dengan spesifikasi di TABEL I.

TABEL I
Lingkungan Komputasi Pelatihan Model

Komputasi	Accelerator	VRAM	Operating System
Google Colaboratory	T4 GPU	16GB	Windows 11
Kaggle Notebooks	2 T4 GPU	32GB	Windows 11
Komputer	RTX 4060	8GB	Ubuntu

Lokal			24.04

1. HASIL KLASIFIKASI TANPA METODE SOLUSI USULAN

Pada tahapan awal, diteliti perbedaan klasifikasi setiap model untuk mendapatkan model *baseline* terbaik sebelum dilakukan pelatihan dengan metode yang diusulkan. Evaluasi model klasifikasi tanpa menggunakan solusi usulan dilakukan pada model-model pretrained dengan hyperparameter terbaik pada masing-masing model yang sudah dicoba, yaitu : ResNet18, ResNet50, ViT Base (86 juta parameter), ViT-Large (307 juta parameter), dan DINO. Untuk setiap model digunakan metriks evaluasi akurasi pada data latih, data validasi, dan data tes. Akurasi tes yang dimaksud pada eksperimen ini adalah nilai akurasi pada data data tes *public* dalam kompetisi Data Mining UNESA 2025 pada platform *kaggle*. Beberapa model tidak memiliki *test accuracy* dikarenakan *training accuracy* dan *validation accuracy* yang buruk sehingga tidak dikumpulkan pada platform *kaggle.*

TABEL II

Hasil Evaluasi Model Tanpa Menggunakan Metode Usulan

Model	*Training Accuracy*	*Validation Accuracy*	*Test Accuracy*
ResNet-18	82.33 %	80.50 %	-
ResNet-50	78.08 %	76.13%	-
ViT-Base	99.88 %	90.64 %	91.24 %
ViT-Large	99.85 %	92.37 %	**92.94%**
Dinov2-Large	**100.00 %**	**93.45 %**	91.97%

Dari hasil evaluasi kelima model di **TABEL II**, terbukti bahwa secara akurasi ViT-Large merupakan model baseline yang paling bagus. Dari antara kelima model *pretrained*, ViT-Large memiliki parameter terbanyak, disusul dengan Dinov2-Large, kemudian ViT-Base dan kemudian model ResNet-50 serta ResNet-18. Dari eksperimen pencarian model *baseline* terbaik bisa ditemukan korelasi positif antara jumlah parameter model dengan tingkat akurasi model. Dengan hasil eksperimen di atas, maka dilakukan tahap eksperimen kedua yaitu melakukan pelatihan ulang model dengan menggunakan metode usulan.

2. HASIL KLASIFIKASI MENGGUNAKAN METODE SOLUSI USULAN

TABEL III

Hasil Evaluasi Model Menggunakan Metode Usulan

Model	*Training Accuracy*	*Validation Accuracy*	*Test Accuracy*
ViT-Large	99.88%	92.24 %	90.26%
Dinov2-Large	**100.00 %**	**93.45 %**	**93.67%**

Hasil Evaluasi dari Tabel II dan Tabel III adanya bukti peningkatan akurasi pada Dinov2-Large. Akurasi tes baseline model ini adalah 91.97%. Setelah penerapan metode usulan, akurasi tes meningkat menjadi 93.67%. Ini merupakan peningkatan absolut sebesar 1.7%, yang membuktikan efektivitas metode usulan untuk arsitektur ini. Di saat yang sama, terjadi penurunan performa pada ViT-Large. Sebagai baseline, model ini mendapatkan akurasi tes tertinggi di 92.94%. Namun, setelah penerapan metode usulan yang sama, performanya turun menjadi 90.26%, atau penurunan sebesar 2.68%.

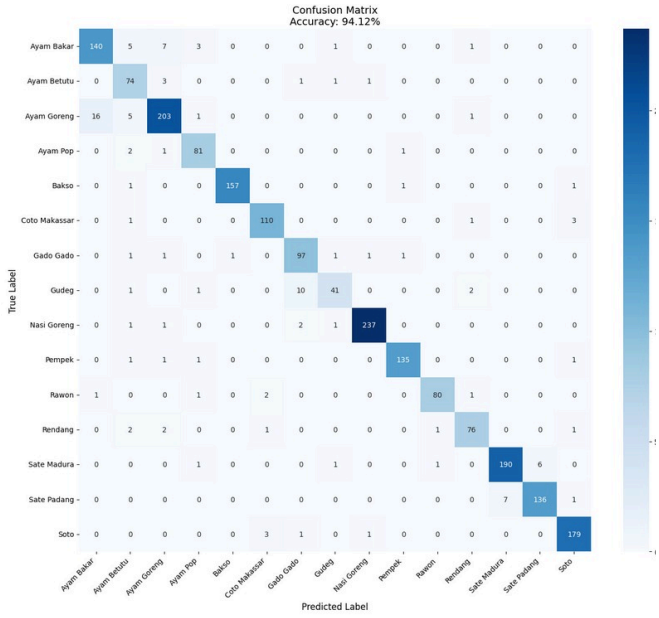
B. ANALISIS PENGUJIAN

Hasil eksperimen yang telah dilakukan mengindikasikan bahwa metode usulan yang mencakup ekstraksi fitur DINOv2, Linear Probing , dan Fine-Tuning dengan Layer-wise Learning Rate Decay (LLRD) terbukti optimal untuk arsitektur DINOv2.

Keberhasilan pada Dinov2-Large menunjukkan bahwa strategi fine-tuning dua tahap (melakukan *freezing* pada backbone terlebih dahulu, kemudian melatih seluruh layer dengan LLRD) sangat efektif untuk mentransfer pengetahuan self-supervised DINO ke tugas klasifikasi makanan spesifik ini. Selain itu, akurasi validasi (93.45%) dan akurasi tes (93.67%) pada model Dinov2-Large usulan sangat berdekatan, menunjukkan bahwa model berhasil melakukan generalisasi dengan baik dan tidak mengalami overfitting pada data validasi.

Sebaliknya, penurunan performa pada ViT-Large mengisyaratkan bahwa strategi fine-tuning yang dioptimalkan untuk DINOv2 tidak cocok untuk arsitektur ViT standar. Penerapan LLRD atau linear probing pada model ViT-Large kemungkinan memerlukan penyesuaian hyperparameter yang berbeda (melalui Optuna) yang mungkin tidak tercakup dalam rentang pencarian yang sama dengan DINOv2.

Untuk mendapatkan gambaran hasil test dan interpretasi akurasi untuk setiap kelas, dataset test dengan jumlah 2057 dilabeli secara manual tanpa teknik semi supervised. Dari hasil model DINOv2 yang sudah dilatih, dilakukan inferensi terhadap dataset test dan dilakukan evaluasi melalui label setiap gambar yang sudah dilakukan secara manual.



Gambar 2. *Confusion Matrix* pada *Test Set* yang sudah dilakukan *labelling* manual

Untuk memahami di mana model terbaik (DinoV2-Large dengan metode usulan) masih melakukan kesalahan, dilakukan analisis *confusion matrix* terhadap hasil prediksi. Gambar 2 menunjukkan confusion matrix dari model DinoV2-Large usulan yang menghasilkan akurasi 94.12% pada test data yang digunakan untuk evaluasi ini. Hasil akurasi 94.12% bukan merupakan representasi dari hasil akurasi pada *submission* di platform kaggle.

Titik kesalahan terbesar model adalah pada klasifikasi Ayam Goreng. Terdapat 16 sampel 'Ayam Goreng' (Label Sebenarnya) yang salah diprediksi sebagai 'Ayam Bakar' (Label Prediksi). Kebingungan ini bersifat timbal balik. 'Ayam Bakar' (Label Sebenarnya) juga keliru diprediksi sebagai 'Ayam Goreng' (7 kali) dan 'Ayam Betutu' (5 kali). Selain itu, 'Ayam Betutu' (Label Sebenarnya) juga 7 kali salah diprediksi sebagai 'Ayam Bakar'. Kesalahan ini bisa diinterpretasikan bahwa model kesulitan membedakan fitur-fitur halus antara hidangan ayam. Secara visual, 'Ayam Bakar' dan 'Ayam Goreng' sama-sama berwarna cokelat dan memiliki presentasi yang serupa, sehingga menjadi tantangan terbesar bagi model.

Di sisi lain, model menunjukkan performa nyaris sempurna untuk kelas-kelas dengan fitur visual yang sangat unik. Kelas seperti Nasi Goreng (237/239 benar), Sate Madura (190/190 benar), Sate Padang (136/137 benar), Pempek (135/137 benar), dan Bakso (157/159 benar) memiliki akurasi yang sangat tinggi. Secara visual, fitur seperti tusuk sate, bentuk 'kapal selam' pada pempek, atau tekstur nasi pada nasi goreng sangat mudah dikenali oleh model dan tidak tumpang tindih dengan kelas lain.

VI. KESIMPULAN

Penelitian ini sukses mencapai tujuannya, yaitu meningkatkan akurasi klasifikasi gambar makanan khas Indonesia melalui penerapan strategi fine-tuning lanjutan pada arsitektur Vision Transformer. Hasil pengujian menunjukkan bahwa metode usulan, yang mengombinasikan ekstraksi fitur DINOv2, Linear Probing, dan Fine-Tuning menggunakan Layer-wise Learning Rate Decay (LLRD), terbukti efektif dan mampu menghasilkan performa terbaik. Model DinoV2-Large yang dilatih dengan metode ini mencapai akurasi tes tertinggi sebesar 93.67%, secara signifikan melampaui performa baseline DinoV2-Large (91.97%) serta baseline terbaik ViT-Large (92.94%).

Temuan penting lainnya adalah bahwa metode fine-tuning yang dioptimalkan untuk DinoV2 tidak cocok untuk arsitektur ViT-Large standar, yang justru mengalami penurunan performa, menegaskan pentingnya menyesuaikan strategi pelatihan dengan spesifikasi arsitektur pre-trained. Meskipun performa keseluruhan sangat tinggi, analisis confusion matrix menunjukkan bahwa tantangan utama model saat ini adalah dalam ranah fine-grained classification, di mana terjadi kebingungan yang signifikan antara kelas-kelas yang secara visual sangat mirip, khususnya antar jenis hidangan ayam. Secara keseluruhan, penelitian ini menyimpulkan bahwa strategi fine-tuning yang cerdas dan disesuaikan adalah kunci untuk mengoptimalkan transfer pengetahuan dari model foundation ke domain klasifikasi makanan spesifik.

Berdasarkan temuan yang ada, terdapat beberapa area yang dapat dieksplorasi dalam penelitian lanjutan untuk meningkatkan kinerja model lebih lanjut. Pertama, disarankan untuk memfokuskan upaya pada teknik Fine-Grained Visual Classification (FGVC) guna mengatasi kesulitan model dalam membedakan kelas yang visualnya mirip, misalnya dengan mengaplikasikan attention mechanisms yang lebih spesifik atau loss functions lanjutan seperti triplet loss. Kedua, penelitian selanjutnya dapat mencakup pengujian dan optimalisasi hyperparameter pada Test-Time Augmentation (TTA) untuk menemukan kombinasi augmentasi yang paling efektif dalam meningkatkan akurasi inferensi. Ketiga, meskipun DinoV2-Large telah memberikan hasil yang unggul, eksplorasi terhadap arsitektur foundation model lain yang lebih baru dan teknik self-supervised learning terkini tetap penting untuk mengetahui apakah ada backbone lain yang dapat memberikan representasi fitur yang lebih superior untuk dataset makanan Indonesia. Penelitian lanjutan ini diharapkan dapat menutup kesenjangan performa pada klasifikasi visual yang subtil.

REFERENSI

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[3] O. Caron et al., "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[4] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington, MA: Morgan Kaufmann, 2017.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," presented at the NAACL, 2019.

[6] Y. Wang, G. Li, and F. Li, "An analysis of test-time augmentation for deep learning," in *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-8.

[7] V. Shankar, R. Roelofs, H. Mania, A. Fang, B. Recht, dan L. Schmidt, "Evaluating Machine Accuracy on ImageNet," Proc. ACM Conf. on Human Factors in Computing Systems (CHI), 2020.

[8] S. Zhang, C. Zhu, J.K.O. Sin, dan P.K.T. Mok, “A Novel Ultrathin Elevated Channel Low-temperature Poly-Si TFT,” *IEEE Electron Device Lett.*, Vol. 20, hal. 569–571, Nov. 1999, doi: 10.1109/55.798046.

[9] M. Wegmuller, J.P. von der Weid, P. Oberson, dan N. Gisin, “Highresolution Fiber Distributed Measurements with Coherent OFDR,” *Proc. ECOC’00*, 2000, paper 11.3.4, hal. 109.

[10] R.E. Sorace, V.S. Reinhardt, dan S.A. Vaughn, “High-speed Digital-to-RF Converter,” U.S. Patent 5 668 842, 16 Sep. 1997.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.