Klasifikasi Citra Makanan Tradisional Indonesia Berbasis SigLIP–*Enhanced* MLP dengan Pendekatan *Self-Training* Adaptif

Solo Bening Nuansa Nanditya[1], Wildan Abid Al Hanif[2], Aqila Khansa Hartanto[3], Winita Sulandari[4]

[1]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret, Surakarta, email: ketuatim@institusi.ac.id
[2]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret, Surakarta, email: anggotatim1@institusi.ac.id
[3]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret, Surakarta, email: aqilakhansahartanto@student.uns.ac.id
[4] Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret, Surakarta, email: winita@mipa.uns.ac.id
Corresponding Author: Solo Bening Nuansa Nanditya

ABSTRAK — Penelitian ini membahas pengenalan otomatis makanan tradisional Indonesia menggunakan pendekatan *semi-supervised learning*. Latar belakang riset berangkat dari menurunnya dokumentasi kuliner lokal serta keterbatasan data berlabel yang menghambat penerapan *deep learning*. Tujuan utama penelitian ini adalah mengembangkan model klasifikasi berbasis SigLIP (*Vision Transformer*) sebagai *feature extractor* dan *Enhanced* MLP sebagai *classifier*, menerapkan *self-training* adaptif berbasis *pseudo-labeling*, serta mengevaluasi kinerja model dengan metrik komprehensif. Dataset bersumber dari Kaggle yang berisi 15 kelas makanan (±4.257 citra latih dan 2.057 citra uji). Sebanyak 599 citra diberi *seed label* manual, kemudian diekstraksi menggunakan SigLIP menghasilkan vektor fitur berdimensi 1.152 dan dinormalisasi menggunakan *standardization* serta L2-normalization. Proses *self-training* iteratif menambahkan 3.305 *pseudo-label* tervalidasi melalui k-NN *cosine similarity*, sehingga total data berlabel meningkat menjadi 3.904 citra yang seimbang antar kelas. Pelatihan akhir menggunakan *Enhanced* MLP dengan blok Residual FFN dan SE-Attention, dioptimasi AdamW dengan *label smoothing* serta *cosine schedule*. Hasil menunjukkan *validation accuracy* sebesar 97,29% dan *Macro F1-Score* 97,27%, menunjukkan performa tinggi dan stabil di semua kelas. Submisi ke Kaggle menghasilkan akurasi publik 93,19%. Kombinasi SigLIP, *Enhanced* MLP, dan *self-training* adaptif terbukti efektif untuk klasifikasi citra pada *dataset* minim label serta relevan bagi upaya digitalisasi kuliner lokal Indonesia.

KATA KUNCI — SigLIP, *Pseudo-Labeling, Enchanced MLP*, Klasifikasi Citra, Makanan Tradisional

I. PENDAHULUAN

A. Latar Belakang

Makanan tradisional merupakan bagian dari warisan budaya yang memiliki nilai historis, sosial, dan ekonomi yang tinggi. Setiap daerah di Indonesia memiliki makanan khas yang mencerminkan identitas dan kearifan lokal masyarakatnya. Namun, seiring dengan pesatnya arus globalisasi dan modernisasi, pengetahuan serta apresiasi masyarakat terhadap makanan tradisional mulai berkurang. Hal ini berpotensi menyebabkan berkurangnya dokumentasi dan pelestarian kuliner lokal, padahal keberagaman makanan tradisional Indonesia memiliki potensi besar untuk dikembangkan dalam bidang pariwisata, ekonomi kreatif, serta industri pangan [1].

Perkembangan teknologi kecerdasan buatan (*artificial intelligence*) khususnya di bidang *computer vision* telah membuka peluang baru dalam pendataan dan pengenalan makanan tradisional. Melalui teknik klasifikasi citra berbasis *deep learning*, sistem dapat mengenali jenis makanan secara otomatis berdasarkan citra visual. Teknologi ini berpotensi digunakan untuk katalogisasi kuliner lokal, sistem rekomendasi makanan, hingga promosi wisata gastronomi [2].

Namun, implementasi sistem klasifikasi citra makanan tradisional Indonesia masih menghadapi beberapa kendala. Pertama, keterbatasan jumlah data berlabel, yang menyebabkan model sulit dilatih dengan baik. Proses pelabelan manual membutuhkan waktu dan tenaga yang besar, sedangkan ketersediaan dataset kuliner lokal masih sangat terbatas. Kedua, variasi kualitas citra yang tinggi, seperti pencahayaan yang tidak merata, latar belakang kompleks, dan resolusi gambar rendah, dapat menurunkan performa model dalam mengenali pola visual [3].

Untuk menjawab tantangan tersebut, penelitian ini mengusulkan pengembangan model klasifikasi citra makanan tradisional berbasis semi-supervised learning dengan memanfaatkan transfer learning SigLIP (*Vision Transformer*) sebagai ekstraktor fitur dan menerapkan *self-training* adaptif (*pseudo-labelling*) untuk memperluas data berlabel secara otomatis. Pendekatan ini diharapkan dapat meningkatkan akurasi klasifikasi meskipun data berlabel terbatas, sekaligus memperbaiki kualitas dataset melalui proses pembersihan citra dan normalisasi fitur.

B. Urgensi Penelitian

Penelitian ini memiliki urgensi yang tinggi karena belum banyak studi yang berfokus pada pengenalan otomatis makanan tradisional Indonesia menggunakan pendekatan semi-supervised. Sebagian besar penelitian terdahulu masih mengandalkan metode *supervised learning* dengan dataset besar dan berlabel lengkap, seperti *Food-101* atau *UEC-Food256*, yang tidak merepresentasikan keanekaragaman kuliner lokal [1].

Melalui pendekatan *pseudo-labelling* adaptif, penelitian ini berpotensi memberikan kontribusi nyata dalam pengembangan sistem digitalisasi kuliner Indonesia yang efisien dan berkelanjutan. Selain itu, penelitian ini juga menjadi langkah awal menuju integrasi teknologi pengenalan citra ke dalam

aplikasi edukasi, promosi wisata, dan pengelolaan basis data kuliner daerah.

C. Tujuan Penelitian

Tujuan yang ingin dicapai melalui penelitian ini adalah sebagai berikut:

1. Mengembangkan model klasifikasi citra makanan tradisional Indonesia berbasis SigLIP (*Vision Transformer*) sebagai ekstraktor fitur dan Enhanced MLP sebagai *classifier*.
2. Menerapkan metode *self-training* adaptif (pseudo-labelling) untuk memanfaatkan data tidak berlabel dalam proses pelatihan model.
3. Mengevaluasi kinerja model dengan metrik evaluasi dengan metode pendekatan yang digunakan.
4. Meningkatkan kualitas dataset melalui proses pembersihan citra (*image cleaning*) dan normalisasi fitur (Z-score dan L2 normalization).

D. Manfaat Penelitian

1. Manfaat Teoritis
   Secara teoritis, penelitian ini diharapkan dapat memperkaya kajian dalam bidang *computer vision* dan *machine learning*, khususnya dalam penerapan metode *semi-supervised* learning untuk klasifikasi citra dengan keterbatasan label. Penelitian ini juga diharapkan dapat menjadi referensi akademik dalam pengembangan model berbasis *Vision Transformer* (SigLIP) di bidang pengenalan kuliner lokal.
2. Manfaat Praktis
   Secara praktis, hasil penelitian ini dapat dimanfaatkan dalam:
   1) Pengembangan sistem pengenalan makanan tradisional otomatis untuk mendukung promosi wisata kuliner Indonesia.
   2) Penyusunan basis data digital makanan tradisional yang terstandar dan mudah diakses oleh masyarakat serta pelaku industri kreatif.
   3) Implementasi pada aplikasi edukasi dan rekomendasi makanan lokal berbasis kecerdasan buatan.

II. STUDI LITERATUR

A. Data Mining

Data mining merupakan proses untuk menemukan dan mengenali berbagai pola dari sekumpulan data. Berdasarkan banyaknya data yang tersimpan dalam suatu *database*, kegiatan ini bertujuan untuk menelusuri kemungkinan adanya keterkaitan, kecenderungan, atau pola tertentu yang dianggap bermanfaat. Hasil dari proses tersebut kemudian dapat dimanfaatkan oleh organisasi atau perusahaan pemilik *database* sebagai dasar pengambilan keputusan maupun pengembangan strategi bisnis [4].

B. SigLIP

Dalam domain pemrosesan bahasa dan visi komputer, pendekatan *pre-training* multimodal telah merevolusi kemampuan model untuk memahami dan menghasilkan representasi yang koheren antara teks dan gambar. SigLIP (*Sigmoid Loss for Language-Image Pre-Training*) [5], menandai terobosan signifikan dengan mengganti fungsi loss kontrasif tradisional seperti InfoNCE dengan fungsi sigmoid yang lebih sederhana dan efisien secara komputasional. Model ini memanfaatkan pasangan teks-gambar yang dihasilkan secara otomatis dari sumber web skala besar, sehingga mencapai performa unggul dalam tugas-tugas *zero-shot* seperti pengambilan gambar-teks dan klasifikasi gambar, sering kali melampaui model seperti CLIP dengan biaya pelatihan yang lebih rendah.

SigLIP 2, sebagai evolusi dari arsitektur pendahulunya, memperluas paradigma ini dengan integrasi teknik adaptif yang lebih canggih untuk menangani heterogenitas data multimodal pada skala yang lebih besar [6]. Secara spesifik, SigLIP 2 mengadopsi mekanisme normalisasi dinamis berbasis sigmoid yang dimodifikasi untuk mengurangi sensitivitas terhadap noise label, yang sering kali menjadi tantangan dalam dataset *web-scraped* yang tidak terkurasi [7]. Inovasi utama terletak pada penggunaan *loss* sigmoid yang dioptimalkan dengan faktor regularisasi adaptif, yang memungkinkan model untuk secara dinamis menyesuaikan bobot antar-modalitas selama fase *pre-training*. Hasil empiris dari [6] menunjukkan bahwa SigLIP 2 mencapai peningkatan akurasi *zero-shot* sebesar 4,7% pada *benchmark* ImageNet dibandingkan SigLIP asli, sambil mengurangi konsumsi memori GPU hingga 25% melalui distilasi pengetahuan yang terintegrasi.

C. Enchanced Multi Layer Perceptron

Enhanced Multi-Layer Perceptron (*Enhanced* MLP) merupakan pengembangan dari MLP klasik yang telah lama menjadi dasar jaringan saraf tiruan karena kemampuannya memodelkan fungsi non-linear melalui arsitektur *feedforward*. Namun, MLP konvensional sering mengalami keterbatasan dalam menangani data kompleks seperti citra dan deret waktu, terutama dari sisi efisiensi komputasi dan kemampuan representasi fitur. Untuk mengatasinya, *Enhanced* MLP mengintegrasikan mekanisme modern seperti *Squeeze-and-Excitation (SE)-Attention Blocks*, *Residual Feed-Forward Layers (FFN)*, serta *Progressive Dimension Reduction* guna meningkatkan performa tanpa mengorbankan kesederhanaan struktur. Pendekatan ini terinspirasi dari *MLP-Mixer* [8] yang menggantikan mekanisme attention dengan MLP murni untuk visi komputer, serta ResMLP [9] yang menekankan pentingnya koneksi residual bagi stabilitas pelatihan.

Komponen utama pertama, *SE-Attention Blocks*, didasarkan pada mekanisme *Squeeze-and-Excitation (SE)* yang diperkenalkan oleh [10]. Mekanisme ini melakukan kompresi global pada tensor fitur melalui *global average pooling*, menghasilkan vektor saluran yang kemudian diproses oleh MLP *bottleneck* untuk menghitung bobot skala adaptif. Bobot tersebut diterapkan kembali pada fitur asli, sehingga saluran yang paling relevan ditekankan secara dinamis. Integrasi SE ke dalam MLP terbukti meningkatkan kemampuan model dalam menangkap hubungan antar-saluran tanpa menambah kompleksitas berlebih, seperti ditunjukkan pada arsitektur hibrid *CNN–Transformer* dan model seperti *FasterMLP* [11] yang menggabungkan *attention* dan *wavelet downsampling*.

Komponen kedua, *Residual FFN Layers*, merupakan adaptasi dari *Feed-Forward Network* yang dilengkapi dengan koneksi residual sebagaimana diperkenalkan pada ResNet dan kemudian diadopsi oleh *Transformer*. Koneksi residual ini mencegah *vanishing gradient* dan membantu pelatihan jaringan yang lebih dalam. Dalam konteks *Enhanced MLP*, lapisan residual memungkinkan pertukaran informasi antar-token atau *patch* secara paralel tanpa bergantung penuh pada *self-attention*. Pendekatan ini juga diterapkan dalam ResMLP [9] dan model

RAM (*Replace Attention with MLP*) untuk tugas klasifikasi citra maupun peramalan deret waktu multivariat.

D. Self-Training

Self-training merupakan paradigma *semi-supervised learning* yang memungkinkan model klasifikasi gambar untuk memperluas dataset pelatihannya secara iteratif dengan memanfaatkan data tidak berlabel yang melimpah. Proses dimulai dengan pelatihan supervised pada himpunan data berlabel kecil, dilanjutkan dengan inferensi terhadap data tidak berlabel, pemilihan *pseudo-label* berdasarkan ambang kepercayaan (*confidence threshold*), dan augmentasi dataset untuk iterasi berikutnya. Dalam konteks visi komputer, strategi ini terbukti efektif membangun dataset pelatihan berskala besar tanpa biaya anotasi tambahan.

Pseudo-labelling sebagai mekanisme inti *self-training* merujuk pada proses penugasan label sementara pada sampel tidak berlabel menggunakan keluaran probabilistik model terlatih [12]. Label ini dapat bersifat keras (*argmax*) atau lunak (soft probabilities), dan kualitasnya dikontrol melalui dua pendekatan utama: (1) *thresholding* tetap atau adaptif terhadap *confidence score*, serta (2) regularisasi konsistensi antara augmentasi lemah dan kuat. Pada tugas klasifikasi gambar, teknik ini telah berhasil meningkatkan akurasi top-1 ImageNet hingga 88,4 % melalui arsitektur *Noisy Student* yang menggabungkan *teacher-student distillation* dengan *pseudo-label* berkualitas tinggi.

Untuk memitigasi *noise pseudo-label*, beberapa penelitian mengusulkan *reliable sample mining* dan *label correction* berbasis *graph propagation* [13]. Pendekatan ini memastikan hanya gambar dengan *pseudo-label* konsisten dan berdistribusi serupa dengan data berlabel asli yang dimasukkan ke dalam dataset pelatihan, sehingga batas keputusan model tetap stabil di wilayah *low-density* [13].

E. Evaluasi Model Klasifikasi

Evaluasi model klasifikasi bertujuan menilai kemampuan model dalam mengklasifikasikan data secara benar. Ukuran dasar yang paling umum digunakan adalah akurasi, yaitu rasio antara jumlah prediksi yang benar dengan total seluruh observasi [14]. Akurasi memberikan gambaran umum mengenai kinerja model secara keseluruhan dan efektif digunakan ketika distribusi kelas relatif seimbang.

Meskipun demikian, akurasi perlu diinterpretasikan dengan hati-hati karena dapat menyesatkan pada data dengan ketidakseimbangan kelas yang tinggi. Oleh karena itu, beberapa literatur menyarankan untuk melengkapi akurasi dengan metrik tambahan seperti precision, recall, atau F1-score guna memperoleh evaluasi yang lebih komprehensif. Namun, bila tujuan utama adalah menilai konsistensi prediksi secara umum, akurasi tetap menjadi metrik utama yang paling representatif [14].

III. SOLUSI DAN USULAN

A. Analisis Permasalahan

Permasalahan utama dalam klasifikasi citra makanan tradisional Indonesia adalah keterbatasan data berlabel yang dapat digunakan untuk melatih model pembelajaran mendalam (*deep learning*). Sebagian besar dataset hanya memiliki label sebagian, atau bahkan tidak memiliki label sama sekali, sehingga model sulit mencapai performa optimal. Selain itu, kualitas citra yang bervariasi terutama akibat perbedaan pencahayaan, sudut pengambilan, dan latar belakang menyebabkan distribusi fitur visual menjadi tidak homogen.

Masalah lain muncul pada biaya anotasi manual yang tinggi, di mana pelabelan ribuan gambar memerlukan tenaga ahli yang memahami jenis makanan tradisional di tiap daerah. Akibatnya, proses pengumpulan dan pelabelan data menjadi lambat dan tidak efisien. Hal ini berdampak pada performa model yang tidak mampu melakukan generalisasi dengan baik terhadap citra baru atau kondisi pencahayaan yang berbeda.

Selain keterbatasan data dan variasi visual, arsitektur model juga menjadi tantangan. Banyak model klasifikasi citra yang memiliki kompleksitas tinggi (seperti ViT atau CLIP), namun tidak efisien secara komputasi jika diterapkan pada perangkat dengan sumber daya terbatas. Oleh karena itu, dibutuhkan solusi yang tidak hanya akurat, tetapi juga efisien dan adaptif terhadap kondisi dataset yang minim label.

B. Konsep Solusi yang Diusulkan

Penelitian ini mengusulkan pengembangan model klasifikasi citra makanan berbasis SigLIP dan Enhanced MLP dengan mekanisme *self-training* adaptif. Pendekatan ini menggabungkan tiga komponen utama:

1. SigLIP sebagai *feature extractor*, yaitu model *Vision Transformer* yang telah dilatih secara multimodal (teks–gambar) dengan fungsi *loss* sigmoid. Model ini berperan mengekstraksi representasi visual yang kaya konteks dari setiap citra makanan.
2. Enhanced MLP sebagai classifier, yang dilengkapi dengan *Squeeze-and-Excitation (SE) Attention* dan *Residual Feed-Forward Network* untuk meningkatkan efisiensi dan stabilitas pelatihan tanpa kompleksitas berlebih.
3. *Self-training* adaptif (*pseudo-labelling*), yang memanfaatkan data tidak berlabel dengan memberikan label semu berdasarkan prediksi model berlabel sebelumnya. Proses ini dilakukan secara iteratif untuk memperluas dataset pelatihan dan memperkuat kemampuan generalisasi model.

Kombinasi ketiga pendekatan ini diharapkan mampu meningkatkan performa klasifikasi meskipun data berlabel terbatas, sekaligus menurunkan kebutuhan sumber daya komputasi dibanding model transformer penuh.

C. Rancangan Sistem yang Diusulkan

Arsitektur sistem secara umum terdiri atas empat tahap utama:

1. Input Data: *Dataset* citra makanan tradisional (berlabel dan tidak berlabel) dari Kaggle menjadi input utama.
2. *Feature Extraction* (SigLIP): Citra diproses menggunakan *SigLIP pre-trained model* untuk menghasilkan vektor fitur berdimensi 1152.
3. *Self-Training Loop*:
   a. Model MLP awal dilatih dengan data berlabel.
   b. Model tersebut melakukan inferensi terhadap data tidak berlabel untuk menghasilkan pseudo-label dengan ambang kepercayaan ≥0,9.
   c. Data berlabel asli + *pseudo-label* berkualitas tinggi digunakan kembali untuk melatih model.
   d. Proses diulang hingga konvergen.
4. *Classification Output*: Model akhir menghasilkan prediksi kelas makanan dengan probabilitas tiap kategori.

D. Keunggulan dan Manfaat Solusi

Berikut keunggulan dan manfaat solusi:

1. Kinerja Tinggi dengan Data Terbatas

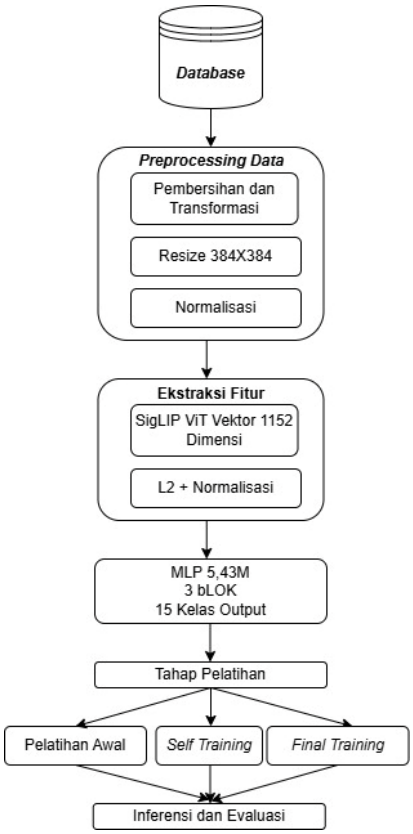
2. Efisiensi Komputasi
3. Generalisasi yang Lebih Baik
4. Mendukung Digitalisasi Kuliner Lokal

E. Usulan Pengembangan Lanjutan

Beberapa pengembangan yang dapat dilakukan antara lain:

1. Integrasi Data Multimodal yaitu Menggabungkan data teks (deskripsi bahan dan daerah asal) untuk meningkatkan konteks klasifikasi.
2. Peningkatan Kualitas Pseudo-Label yaitu Menggunakan mekanisme *teacher–student distillation* agar pseudo-label lebih stabil.
3. Implementasi Aplikasi Nyata yaitu Mengembangkan aplikasi mobile berbasis AI yang dapat mengenali makanan tradisional secara real-time.

IV. METODE PENELITIAN



Gambar 1. Alur penelitian

A. Dataset

Dataset yang digunakan dalam penelitian ini bersumber dari Kaggle, berupa citra digital 15 jenis makanan tradisional Indonesia yang terdiri atas 4.257 gambar latih dan 2.057 gambar uji. Meskipun terdapat 15 kelas makanan tradisional, data latih disediakan tanpa label eksplisit baik dalam bentuk subfolder maupun petunjuk pada nama file, sehingga pelabelan dilakukan menggunaka pendekatan-pendekatan yang memanfaatkan *domain knowledge* dari peneliti dan pendekatan secara algoritmik.

B. Praproses Data

Praproses data dilakukan dalam dua tahap utama, yaitu pembersihan gambar dan transformasi data. Pada tahap pembersihan, gambar difilter untuk menghilangkan *noise* dengan tiga kriteria: (1) rasio piksel putih >98%, (2) standar deviasi intensitas piksel <3,0, dan (3) rata-rata gradien tepi <0,5. Gambar yang lolos filter kemudian di-*resize* menjadi ukuran 384×384 piksel menggunakan metode interpolasi *Lanczos* serta dikonversi ke ruang warna RGB. Selanjutnya, pada tahap transformasi, setiap gambar dikonversi menjadi tensor PyTorch, dinormalisasi menggunakan nilai mean dan standard deviation [0,5, 0,5, 0,5] untuk setiap kanal warna, serta dilengkapi augmentasi data standar (*random horizontal flip* dan *random rotation* ±10°) guna meningkatkan robustitas model. *Dataset* yang telah dipraroses kemudian dibagi menjadi data latih dan data validasi dengan rasio 85:15 menggunakan *stratified sampling* untuk menjaga distribusi kelas.

C. Ekstraksi Fitur

Ekstraksi fitur visual dilakukan menggunakan model jaringan saraf konvolusional SigLIP (varian vit_so400m_patch14_siglip_384), yang diakses melalui *library* timm dan telah dilatih sebelumnya (*pre-trained*). Model dijalankan dalam mode inferensi (*evaluation mode*). Setelah melewati lapisan model, *Global Average Pooling* diterapkan pada peta fitur untuk mereduksi dimensi spasial, menghasilkan vektor fitur berdimensi 1152 untuk setiap gambar yang telah diprapemrosesan. Untuk efisiensi komputasi, proses ekstraksi dilaksanakan dengan ukuran *batch* 64 dan memanfaatkan akselerasi GPU. Semua vektor fitur hasil ekstraksi kemudian di-*cache* dalam format NPZ untuk reproduksibilitas.

D. Normalisasi Fitur (Standardisasi dan L2-Normalization)

Setelah ekstraksi, vektor fitur dinormalisasi menggunakan mean dan standar deviasi yang dihitung dari seluruh data pelatihan untuk menjaga konsistensi skala antar fitur. Selanjutnya, normalisasi L2 diterapkan pada setiap vektor fitur sehingga panjang vektor menjadi 1 (*unit norm*), yang bertujuan meningkatkan stabilitas dan konvergensi model berbasis metrik jarak seperti *cosine similarity* atau klasifikasi berbasis prototipe. Proses ini dilakukan secara terpisah untuk data latih dan data uji, namun menggunakan statistik normalisasi yang sama (dari data latih) agar tidak terjadi kebocoran informasi.

E. Arsitektur Model

Arsitektur model yang diusulkan merupakan Improved Lightweight MLP yang dirancang untuk klasifikasi 15 kelas makanan. Model ini terdiri dari proyeksi awal, tiga blok utama berbasis *Squeeze-and-Excitation* (SE) *attention* dan Residual *Feed-Forward Network* (FFN), serta head klasifikasi linier. Detail arsitektur ditunjukkan pada Tabel 9.

Tabel 1. Arsitektur model

Layer/Blok	Deskripsi
Input	1152, L2 + Normalisasi
Input Projection	Linear(1152, 768) + LN+ GELU + Drop(0.15)
Blok 1	
Squeeze-and-Excitation 1	SE (r=8)
Residual FFN 1	LN + Linear(expand 2.2×)+ GELU+ Drop + Linear(768, 768) + residual
Transition 1	LN + GELU + Drop(0.125) + Linear(768, 512)
Blok 2	
Squeeze-and-Excitation 2	SE (r=8)
Residual FFN 2	LN + Linear(expand 2.0×) + GELU+ Dropout + Linear(512, 512) + residual

Transition 2	LN + GELU + Drop(0.10) + Linear(512, 256)
Blok 3	
Squeeze-and-Excitation 3	SE (r=8)
Residual FFN 3	LN + Linear(expand 1.8×) + GELU+ Drop + Linear(256, 256) + residual
LayerNorm	Final LayerNorm
Classification Head	Linear(256, 15)

Arsitektur model *Enchanced MLP* dengan tiga blok Residual FFN bertingkat dan mekanisme Squeeze-and-Excitation (r=8) untuk *channel* attention dengan total parameter ~5,43 juta.

F. Metode Pelatihan Model

Pelatihan model dilakukan menggunakan pendekatan *self-training* dengan arsitektur *Enchanced MLP* yang telah dijelaskan pada subbagian sebelumnya. Fokus utama tahap ini adalah mengoptimalkan kemampuan generalisasi model melalui kombinasi pelatihan terawasi awal (*initial supervised training*) dan pelatihan lanjutan berbasis data tidak berlabel. Pendekatan ini memungkinkan model secara adaptif memperluas pengetahuan dari data berlabel terbatas tanpa memerlukan peningkatan kompleksitas arsitektur maupun sumber daya komputasi yang besar.

1). Pelatihan Awal (*Initial Supervised Training*)

Tahap pertama menggunakan subset data berlabel. Oleh karena itu model dapat dilatih dengan pendekatan *supervised learning*.

Tabel 2. *Initial supervised training*

Komponen	Konfigurasi
Optimizer	AdamW
Learning Rate	6×10^{-4}
Weight Decay	1×10^{-4}
Loss Function	CrossEntropy + Label Smoothing 0.1
Learning Rate Scheduler	Warmup 10% + Cosine Annealing
Batch Size	64
Epoch	20

2). *Self-Training*

Model awal kemudian diperluas memalui lima ronde self-training iteratif menggunakan citra tidak berlabel. Pada setiap ronde:

1. Model menghasilkan prediksi dan skor kepercayaan untuk seluruh data tidak berlabel.
2. Data pseudo-label dipilih menggunakan ambang kepercayaan adaptif per kelas (0,90–0,75) dengan maksimum 350 sampel per kelas.
3. Validasi dilakukan menggunakan SimilarityValidator berbasis k-NN *cosine similarity* dengan k = 7. Suatu pseudo-label diterima apabila memenuhi salah satu dari dua kondisi, yaitu: (1) setidaknya 60% tetangga terdekat memiliki label yang sama (*strong agreement*), atau (2) 40–59% tetangga memiliki kesamaan label namun disertai tingkat kepercayaan (*confidence*) ≥ 0,85 (*moderate agreement*).
4. Data yang gagal validasi dihapus dari pool untuk mencegah *noise*, dan model diretrain dengan data gabungan (seed + pseudo-label tervalidasi) disertai regularisasi Gaussian noise (σ=0,005).

3). *Final Training*

Setelah proses *self-training* selesai, dataset latih telah berkembang menjadi ribuan sampel berkualitas tinggi yang terdiri atas data seed dan *pseudo-label* tervalidasi. Pelatihan akhir menggunakan konfigurasi model yang sama dengan 55 *epochs* diperkuat dengan *class weighting* berdasarkan distribusi data terbaru, serta augmentasi fitur berupa penambahan *Gaussian noise* kecil untuk meningkatkan *robustness*.

G. Test Time Augmentation

Untuk meningkatkan stabilitas dan ketahanan model saat inferensi, digunakan strategi *Test-Time Augmentation (TTA)*. Berbeda dari augmentasi pelatihan, TTA menghasilkan beberapa versi representasi fitur dari data uji dan menggabungkannya menjadi satu prediksi akhir yang lebih andal.

Penelitian ini menerapkan 12 augmentasi melalui empat strategi utama. Hasil prediksi tiap augmentasi digabung menggunakan ensemble berbobot berdasarkan confidence rata-rata (*fungsi softmax*), dengan prediksi asli sebagai acuan. Seluruh inferensi dilakukan secara *batch* pada data yang telah dinormalisasi [15].

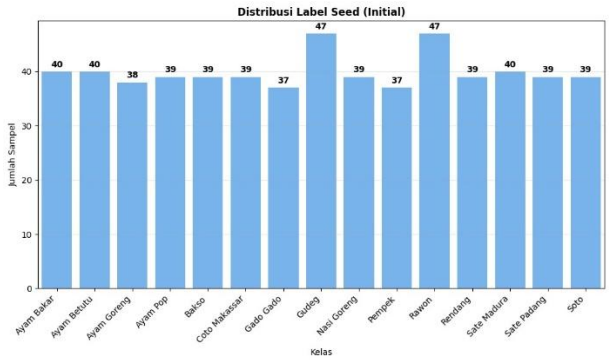
Tabel 3. Strategi augmentasi

Strategi	Deskripsi	Parameter
Feature Dropout	Menghilangkan sebagian fitur secara acak untuk meningkatkan *robustness*	0.10
Adaptive Noise	Menambahkan noise proporsional terhadap skala fitur	0.03
Feature Mixing	Menggabungkan fitur antar-sampel untuk regularisasi	α = 0.15
Channel Shuffle	Mengacak urutan fitur antar grup kecil	1/8 dimensi

V. HASIL DAN ANALISIS PENGUJIAN

A. Pelabelan Manual

Pada tahap awal, sebanyak 599 citra makanan Indonesia dipilih secara acak dan diberi label manual berdasarkan pedoman kelas yang telah ditentukan, menghasilkan *seed labels* dengan distribusi awal yang relatif merata. *Seed labels* ini disimpan dalam file csv dengan format *file_name* dan *label*. Setiap kelas, mulai dari Ayam Bakar hingga Soto, mendapatkan 35–47 citra sehingga tidak ada kelas yang memiliki jumlah yang sangat sedikit. Distribusi label yang merata ini ditunjukkan pada Gambar 1, yang memperlihatkan keseimbangan jumlah sampel antar kelas.



Gambar 1. Distribusi pelabelan manual

Gambar 2 menampilkan contoh citra berlabel seperti rendang dengan kuah kental khas, sate lilit yang digulung pada

batang serai, dan gulai tunjang dengan potongan tulang sumsum, mencerminkan keragaman visual dalam setiap kelas. *Seed* dataset ini menjadi fondasi kuat bagi ekstraksi fitur SigLIP dan pelatihan model *Enhanced MLP*.

Gambar 2. Sampel berlabel

B. Praproses Data

Pada tahap ini dilakukan pembersihan citra mentah untuk menghilangkan gambar yang tidak layak digunakan dalam pelatihan. Dari total 4.257 gambar dalam data latih awal, sebanyak 4.097 citra (96,2%) berhasil dipertahankan setelah melalui tahap pembersihan, sementara 160 citra (3,8%) dieliminasi karena terdeteksi sebagai noise. Citra yang dieliminasi umumnya memiliki karakteristik proporsi piksel putih lebih dari 98%. Sampel citra yang dihapus dapat dilihat pada Gambar 3.

Gambar 3. Dataset yang dihapus

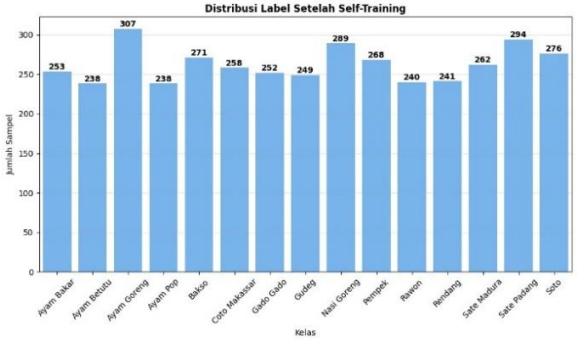
C. Ekstraksi Fitur

Seluruh citra yang telah dibersihkan diekstraksi menggunakan SigLIP ViT SO400M Patch14 sebagai *feature extractor*, menghasilkan vektor fitur berdimensi 1.152 untuk setiap citra. Proses ini dijalankan dalam mode inferensi dengan *mixed precision* sehingga lebih efisien. Seluruh fitur kemudian dinormalisasi melalui standardisasi dan *L2-normalization*, menghasilkan representasi yang stabil, seragam, dan siap digunakan untuk pelatihan MLP serta proses *pseudo-labeling* berbasis kesamaan fitur.

D. Self-training

Model awal, dilatih pada 599 citra berlabel resmi menggunakan *Enchanced MLP* dengan SigLIP sebagai *feature extractor*, mencapai akurasi validasi 93,33% dan akurasi seed 98,50%. *Early stopping* memastikan konvergensi optimal tanpa *overfitting*, menjadikan model ini *baseline* kuat untuk klasifikasi citra makanan Indonesia.

Proses *self-training* adaptif dilaksanakan dalam lima ronde dengan strategi *threshold* dinamis dan pembatasan maksimal per kelas untuk menjaga keseimbangan distribusi. Pada ronde pertama, model menghasilkan 2.376 *pseudo-label* berkualitas tinggi (threshold ≥ 0,92), dan pelatihan ulang dengan data gabungan meningkatkan akurasi validasi menjadi 98,88%. Ronde kedua menurunkan *threshold* secara adaptif dan menambahkan 753 *pseudo-label*, sehingga memperkaya representasi kelas-kelas dengan sampel terbatas. Ronde ketiga melengkapi iterasi dengan 176 pseudo-label tambahan pada threshold yang lebih rendah namun tetap terkontrol, dengan *early stopping self-training* berhenti pada ronde ketiga. Total pseudo-label yang berhasil ditambahkan adalah 3.305 sampel, sehingga jumlah data berlabel meningkat dari 599 menjadi 3.904 citra.. Gambar menunjukkan distribusin kelas setelah *self-training*.



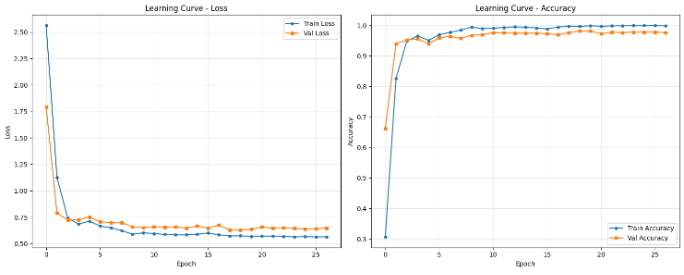
Gambar 4. Distribusi label *setelah self-training*

Distribusi label setelah *self-training* menunjukkan keseimbangan yang baik antar 15 kelas makanan Indonesia, dengan jumlah sampel per kelas berkisar 238–307 citra. Kelas terbanyak adalah Ayam Bakar (307), diikuti Soto (276), sementara kelas tersedikit adalah Ayam Betutu (238). Sampel citra setelah *self-training* dapat dilihat pada Gambar 5.

Gambar 5. Sampel gambar *pseudo-labeled*

E. Final Training

Pelatihan akhir dilaksanakan menggunakan gabungan data berlabel dan *pseudo-label* dengan pemantauan *early stopping*. Gambar 5 menunjukkan kurva *loss* dan *accuracy* pada data latih dan validasi, memperlihatkan pola pembelajaran yang stabil dan konsisten. *Training loss* turun tajam dari sekitar 2,5 menjadi 0,6 pada fase awal, kemudian mendatar pada rentang 0,55–0,60 hingga akhir pelatihan. *Validation loss* mengikuti pola serupa tanpa lonjakan signifikan, dengan gap minimal antara *training* dan *validation loss*, mengindikasikan tidak terjadi *overfitting* meskipun dataset diperbesar melalui *pseudo-labeling*.



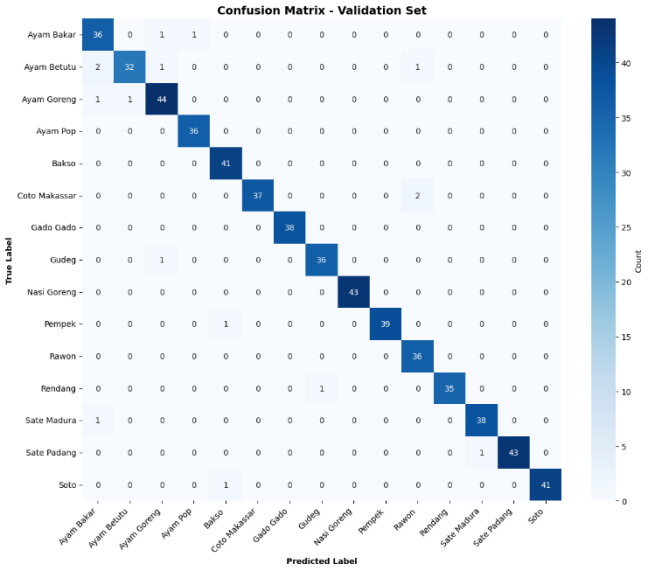
Gambar 6. *Learning curve*

Akurasi meningkat sangat cepat pada fase awal, dari 0,30 menjadi lebih dari 0,90. *Training accuracy* mencapai hampir 1,00 (*perfect training*), sementara *validation accuracy* stabil pada rentang 0,96–0,98 tanpa penurunan tajam (*accuracy drop*) yang biasanya menandakan *overfitting* berat. Stabilitas kurva learning menunjukkan bahwa kombinasi teknik regularisasi seperti *label smoothing*, *dropout*, dan *weight decay*, bersama strategi pelatihan yang digunakan, berhasil mencegah *overfitting* meskipun training accuracy mendekati sempurna. Pola konvergensi cepat pada fase awal juga mengonfirmasi bahwa representasi fitur SigLIP memberikan *starting point* yang sangat baik, sehingga proses pembelajaran berlangsung efisien dengan generalisasi kuat terhadap data validasi.

F. Evaluasi Model

Model akhir dievaluasi menggunakan data validasi untuk mengukur performa klasifikasi secara komprehensif. Hasil

evaluasi menunjukkan akurasi validasi sebesar 0,9729 atau 97,29% dengan Macro F1-Score 0,9727 atau 97,27%, mengindikasikan performa yang sangat baik dan seimbang antar kelas. Kedekatan nilai akurasi dan F1-Score menunjukkan bahwa model tidak hanya akurat secara keseluruhan, tetapi juga konsisten dalam mengenali setiap kelas tanpa bias signifikan terhadap kelas mayoritas.



Gambar 7. *Confusion matrix*

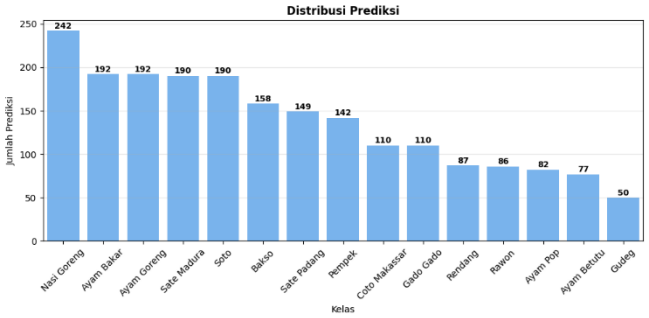
Confusion matrix pada Gambar 7 memperlihatkan detail performa klasifikasi untuk setiap kelas pada validation set. Mayoritas kelas menunjukkan akurasi tinggi dengan nilai diagonal yang dominan, seperti Ayam Goreng (44/46), Nasi Goreng (43/43), Sate Padang (43/44), Soto (41/42), dan Bakso (41/41) yang mencapai akurasi mendekati sempurna. Kelas lain seperti Gado Gado (38/38), Pempek (39/40), Sate Madura (38/39), Coto Makassar (37/39), Ayam Pop (36/36), Ayam Bakar (36/38), Gudeg (36/37), dan Rawon (36/36) juga menunjukkan performa sangat baik dengan tingkat kesalahan minimal.

Beberapa kesalahan klasifikasi terjadi pada kelas dengan karakteristik visual yang mirip. Ayam Betutu mengalami 2 kesalahan prediksi ke Ayam Bakar dan 1 ke Rendang, kemungkinan karena kesamaan tampilan ayam berbumbu dengan warna kecoklatan. Coto Makassar memiliki 2 kesalahan ke Rawon, yang dapat dijelaskan oleh kesamaan visual kuah berwarna gelap pada kedua makanan berkuah tersebut. Rendang mengalami 1 kesalahan ke Gudeg, Pempek ke Bakso, Sate Padang ke Rendang, dan Soto ke Bakso, dengan total kesalahan yang sangat minimal. *Confusion matrix* ini mengonfirmasi bahwa model mampu membedakan mayoritas kelas dengan sangat baik, dengan kesalahan hanya terjadi pada pasangan kelas yang memiliki kesamaan visual signifikan

H. Inferensi dan Submisi

Tahap inferensi dilakukan pada data uji menggunakan model terbaik dengan penerapan *Test-Time Augmentation* (TTA) untuk meningkatkan *robustness* prediksi. TTA menghasilkan beberapa variasi augmentasi dari setiap citra uji, kemudian probabilitas prediksi diagregasi untuk menghasilkan prediksi final yang lebih stabil dan akurat.

Distribusi hasil klasifikasi pada data uji menunjukkan pola yang bervariasi antar kelas, dengan jumlah prediksi berkisar 50–242 citra. Gambar 8 memperlihatkan Nasi Goreng sebagai kelas dengan prediksi terbanyak (242 citra), diikuti Ayam Goreng (194 citra) dan Sate Madura (192 citra), sementara Gudeg memiliki prediksi tersedikit (50 citra). Kelas lainnya terdistribusi dengan Soto (189), Ayam Bakar (187), Bakso (158), Sate Padang (149), Pempek (142), Coto Makassar (111), Gado Gado (110), Rendang (88), Rawon (86), Ayam Pop (81), dan Ayam Betutu (78). Variasi distribusi ini mencerminkan tingkat kesulitan klasifikasi yang berbeda antar kelas berdasarkan karakteristik visual data uji.



Gambar 8. Distribusi prediksi

Hasil submisi ke *platform* Kaggle mencapai skor akurasi publik sebesar 0,93187 atau 93,19%, memvalidasi efektivitas pendekatan self-training adaptif dengan kombinasi SigLIP sebagai *feature extractor*, arsitektur EnhancedMLP, dan strategi *pseudo-labeling threshold* dinamis dalam mengenali makanan khas Indonesia meskipun dimulai dengan data berlabel terbatas.

VI. KESIMPULAN

Dalam penelitian ini, kami berhasil mengembangkan solusi klasifikasi citra makanan tradisional Indonesia berbasis *single-model self-training* tanpa *ensemble*, dengan parameter model hanya ~5.43M, dan dapat mencapai performa kompetitif. Poin-poin utama yang berhasil dicapai:

1. Model yang diusulkan berhasil mengimplementasikan kombinasi SigLIP sebagai *feature extractor* dan *Enhanced* MLP sebagai *classifier* dalam kerangka *semi-supervised* learning berbasis *self-training* adaptif.
2. Proses *self-training* dengan *dynamic confidence threshold* (0,90–0,75) dan validasi berbasis kesamaan fitur mampu memperluas data berlabel dari 599 menjadi 3.904 citra tanpa intervensi anotasi manual.
3. Kinerja model mencapai akurasi validasi sebesar 97,29% dan *Macro F1-Score* sebesar 97,27%, menunjukkan kemampuan klasifikasi yang tinggi dan seimbang antar kelas.
4. Peningkatan kualitas data pelatihan dicapai melalui proses *image cleaning*, *standardization*, L2-normalization, serta strategi Test-Time Augmentation (TTA) yang meningkatkan robustnes dan stabilitas prediksi.
5. Pendekatan SigLIP–*Enhanced* MLP berbasis *self-training* adaptif terbukti efektif, efisien, dan berpotensi diaplikasikan untuk pengenalan makanan tradisional Indonesia dalam konteks digitalisasi kuliner, edukasi, serta sistem rekomendasi berbasis kecerdasan buatan.

REFERENSI

[1] A. Wibisono, H. A. Wisesa, Z. P. Rahmadhani, P. K. Fahira, dan P. Mursanto, "Traditional food knowledge of Indonesia : a new high - quality food dataset and automatic recognition system," *J. Big Data*, 2020, doi: 10.1186/s40537-020-00342-5.

[2] A. Haris, J. Muliadi, R. Yohanes, dan V. Hasbi, "Identification of Indonesian Traditional Foods Using Machine Learning and Supported by Segmentation Methods," vol. 8, no. December, 2024.

[3] M. Y. Kardawi, F. M. Saragih, L. Rahadianti, dan A. M. Arymurthy, "Indonesian Food Classification Using Deep Feature Extraction and Ensemble Learning for Dietary Assessment," vol. 9, no. 5, hal. 2009–2018, 2025.

[4] M. I. Alhafiz, Wirasno, dan A. Solichin, "Segmentasi dengan Metode Active Countour untuk Peningkatan Akurasi Klasifikasi Citra USG Kanjer Payudara Menggunnakan K-Nearest Neighbor(KNN)," vol. 10, no. 1, hal. 34–48, 2025, doi: https://doi.org/10.29100/jipi.v10i1.5681.

[5] X. Zhai dan G. Deepmind, "Sigmoid Loss for Language Image Pre-Training," hal. 1–17.

[6] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, dan I. Alabdulmohsin, "SigLIP 2 : Multilingual Vision-Language Encoders with Improved Semantic Understanding , Localization , and Dense Features," no. February, hal. 1–20, 2025.

[7] M. C. Kenton, L. Kristina, dan J. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," no. Mlm, 1953.

[8] I. Tolstikhin *et al.*, "MLP-Mixer : An all-MLP Architecture for Vision," hal. 1–16.

[9] H. Touvron, M. Cord, A. Joulin, P. Bojanowski, M. Caron, dan A. El-nouby, "ResMLP : Feedforward networks for image classification with data-efficient training".

[10] J. Hu, L. Shen, S. Albanie, G. Sun, dan E. Wu, "Squeeze-and-Excitation Networks," hal. 1–13.

[11] C. Ma, X. Liu, Y. Cao, dan J. Rong, "FasterMLP efficient vision networks combining attention mechanisms and wavelet downsampling," hal. 1–14, 2025.

[12] D. Lee, "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks," no. August, 2015.

[13] S. Learning, O. Chapelle, B. Schölkopf, A. Z. Review, R. P. Thomas, dan C. De Recherche, "Semi-Supervised Learning edited by O . Chapelle , B . Schölkopf and A . Zien . (London : The MIT press , 2006 , 508 pages , hardbound , ISBN 978-0-262-03358-9).," no. September, hal. 4–6, 2014.

[14] I. Technology dan I. Technology, "EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS," vol. 5, no. 2, hal. 1–11, 2015.

[15] H. Zhang, M. Cisse, Y. N. Dauphin, dan D. Lopez-paz, "mixup : B," hal. 1–13, 2018.

LAMPIRAN

Lampiran 1. Link Google Drive File Penyisihan
https://drive.google.com/file/d/1_rnhia1ux2bw7bQq8KzQ2iG9_798E2Iz/view?usp=sharing