

Identifikasi Tanaman Obat Indonesia dengan *Vision Transformer* dan Augmentasi Adaptif

Novi Tristanti^{1*}, Sunardi², Murinto³

¹Program Studi Doktor Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

²Program Studi Teknik Elektro, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

³Program Studi Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

12536083040@webmail.uad.ac.id, sunardi@uad.ac.id, murintokusno@tif.uad.ac.id

Abstract

Manual identification of medicinal plants faces serious challenges due to morphological similarities between species, variations in lighting, and limited availability of botanists in the field. This study proposes an identification system for 100 types of Indonesian medicinal plants using the Vision Transformer (ViT) architecture with a stepwise fine-tuning approach and adaptive image augmentation. The model used is *vit_base_patch16_224* initialized with *ImageNet-1k* pretrained weights, equipped with a classifier head consisting of a series of *LayerNorm*, *Dropout* (0.3), and *Linear* (768→100). The training strategy integrates stepwise freezing (*freeze-unfreeze*) in the first three epochs, the *AdamW* optimizer with a weight decay of 0.05, label smoothing ($\epsilon=0.1$), and cosine-based learning rate scheduling to ensure stable convergence on medium-scale datasets. The dataset used consists of 10,000 images divided using a non-stratified random split with a ratio of 70% training data, 15% validation data, and 15% test data. The evaluation results show that the model achieved an accuracy of 97.3% on the test data, with a macro precision of 0.974, a macro recall of 0.975, and a macro F1-score of 0.973. The macro values were calculated by summing the metric values for each class separately, then dividing by the number of classes without considering the number of samples in each class. The training process lasted for 16 epochs before being terminated by the early stopping mechanism with the best validation accuracy of 97.53% at the 11th epoch. These results demonstrate that the ViT stepwise fine-tuning approach is able to address the challenges of multi-class scale classification on medium-sized datasets effectively and reproducibly.

Keywords: adaptive augmentation, fine-tuning, medicinal plants, vision transformer, image classification

Abstrak

Identifikasi tanaman obat secara manual menghadapi tantangan serius akibat kemiripan morfologi antar spesies, variasi pencahayaan, dan keterbatasan tenaga ahli botani di lapangan. Penelitian ini mengusulkan sistem identifikasi 100 jenis tanaman obat Indonesia menggunakan arsitektur *Vision Transformer* (ViT) dengan pendekatan fine-tuning bertahap dan augmentasi citra adaptif. Model yang digunakan adalah *vit_base_patch16_224* yang diinisialisasi dengan bobot pralatih *ImageNet-1k*, dilengkapi *classifier head* berupa rangkaian *LayerNorm*, *Dropout*(0,3), dan *Linear* (768→100). Strategi pelatihan mengintegrasikan pembekuan bertahap (*freeze-unfreeze*) pada tiga *epoch* pertama, optimizer *AdamW* dengan *weight decay* 0,05, label *smoothing* ($\epsilon=0,1$), dan penjadwalan *learning rate* berbasis kosinus untuk menjamin stabilitas konvergensi pada dataset berskala menengah. Dataset yang digunakan terdiri dari 10.000 citra yang dibagi menggunakan non-stratified random split dengan rasio 70% data latih, 15% data validasi, dan 15% data uji. Hasil evaluasi menunjukkan bahwa model mencapai akurasi sebesar 97,3% pada data uji, dengan *macro precision* 0,974, *macro recall* 0,975, dan *macro F1-score* 0,973. Nilai *macro* dihitung dengan cara menjumlahkan nilai metrik dari setiap kelas secara terpisah, kemudian dibagi dengan jumlah kelas tanpa mempertimbangkan banyaknya sampel pada masing-masing kelas. Proses pelatihan berlangsung selama 16 *epoch* sebelum dihentikan oleh mekanisme *early stopping* dengan akurasi validasi terbaik sebesar 97,53% pada *epoch* ke-11. Hasil ini menunjukkan bahwa pendekatan fine-tuning ViT mampu mengatasi tantangan klasifikasi skala banyak kelas pada dataset berukuran menengah secara efektif dan dapat direproduksi.

Kata kunci: augmentasi adaptif, fine-tuning, tanaman obat, *visionTransformer*, klasifikasi citra

©This work is licensed under a Creative Commons Attribution -ShareAlike 4.0 International License

1. Pendahuluan

Indonesia merupakan negara yang memiliki keanekaragaman hayati sangat tinggi, termasuk ribuan spesies tanaman obat dan herbal yang telah dimanfaatkan secara turun-temurun dalam sistem pengobatan tradisional maupun pengembangan farmasi modern. Berdasarkan data keanekaragaman tumbuhan nasional, terdapat lebih dari dua ribu jenis tanaman herbal yang tersebar di seluruh nusantara, dengan potensi ekonomi, ekologi, dan kesehatan yang signifikan. Namun, identifikasi tanaman obat secara

manual masih menjadi tantangan utama bagi masyarakat umum, petani, hingga praktisi kesehatan tradisional [1], [2]. Banyak spesies memiliki kemiripan morfologi tinggi pada bagian daun, batang, atau bunga, sehingga rentan terhadap kesalahan identifikasi yang dapat berdampak pada keamanan penggunaan herbal.

Perkembangan kecerdasan buatan, khususnya dalam bidang *computer vision*, telah membuka peluang untuk mengotomatisasi proses identifikasi tanaman berbasis citra visual [3]. Selama beberapa tahun terakhir, *Convolutional Neural Network* (CNN) menjadi

arsitektur dominan dalam tugas klasifikasi citra botani [4]. Meskipun efektif dalam mengekstraksi ciri lokal, CNN memiliki keterbatasan dalam menangkap dependensi spasial jangka panjang dan konteks global antarbagian tanaman. Kelemahan ini menjadi kritis ketika mengklasifikasikan objek dengan variasi latar belakang, pencahayaan, atau oklusi parsial yang sering ditemui pada citra tanaman di kondisi lapang [5], [6].

Vision Transformer (ViT) hadir sebagai alternatif yang memperlakukan citra sebagai sekuens *patch* dan memanfaatkan mekanisme *self-attention* untuk memodelkan hubungan global secara eksplisit. Studi terkini menunjukkan bahwa ViT mampu melampaui performa CNN pada *dataset* berskala besar karena kemampuannya memahami pola visual secara holistik [7], [8]. Namun, penerapan ViT pada *dataset* dengan skala menengah (sekitar 10.000 citra) menghadapi tantangan serius berupa risiko *overfitting*, ketidakstabilan pelatihan, dan hilangnya pengetahuan pralatih (*catastrophic forgetting*). Oleh karena itu, diperlukan strategi pelatihan yang dirancang khusus agar model dapat beradaptasi dengan domain spesifik tanpa mengorbankan kemampuan generalisasi.

Penelitian ini mengusulkan pengembangan sistem identifikasi 100 jenis tanaman obat Indonesia menggunakan arsitektur *Vision Transformer* (ViT) dengan pendekatan *stepwise fine-tuning* dan augmentasi adaptif. Kebaruan penelitian ini terletak pada pengusulan dan evaluasi suatu *pipeline* pelatihan yang mengintegrasikan strategi *stepwise fine-tuning* (*freeze-unfreeze*), *label smoothing*, *cosine learning rate scheduling*, dan augmentasi adaptif untuk identifikasi 100 jenis tanaman obat Indonesia menggunakan *Vision Transformer*. Penelitian ini berfokus pada evaluasi performa dari integrasi strategi tersebut sebagai satu kesatuan *pipeline*, bukan pada pembuktian kontribusi masing-masing komponen secara terpisah. Pada penelitian ini, istilah augmentasi adaptif mengacu pada strategi pemilihan dan penyusunan *pipeline* augmentasi yang disesuaikan dengan karakteristik morfologi tanaman obat Indonesia. Adaptivitas yang dimaksud bukan berupa penyesuaian parameter augmentasi secara otomatis selama proses pelatihan (*dynamic parameter adaptation*), melainkan penyesuaian jenis transformasi berdasarkan karakteristik objek penelitian. Oleh karena itu, transformasi seperti *RandomResizedCrop*, *ColorJitter*, *RandomHorizontalFlip*, dan *RandAugment* ($n = 2$, $m = 9$) dipilih untuk meningkatkan keragaman data latih tanpa menghilangkan ciri diagnostik tanaman, sedangkan *VerticalFlip* tidak digunakan karena dapat menghasilkan orientasi daun yang tidak sesuai dengan kondisi alami. Dengan demikian, augmentasi adaptif pada penelitian ini merupakan pendekatan yang menerapkan prinsip *domain-specific augmentation*, yaitu menyesuaikan pemilihan teknik augmentasi dengan karakteristik morfologi tanaman agar informasi visual penting tetap terjaga sekaligus meningkatkan kemampuan generalisasi model [9].

Tujuan penelitian ini adalah: (1) membangun *pipeline*

pengolahan citra dan pelatihan ViT yang stabil dan dapat direproduksi untuk *dataset* 100 kelas tanaman obat; (2) mengevaluasi performa model menggunakan metrik akurasi, presisi, *recall*, dan *F1-score*. Hasil penelitian ini diharapkan dapat menjadi fondasi sistem identifikasi digital yang akurat, efisien, dan siap diintegrasikan ke dalam aplikasi pendukung pelestarian biodiversitas, edukasi kesehatan herbal, maupun sistem informasi botani nasional.

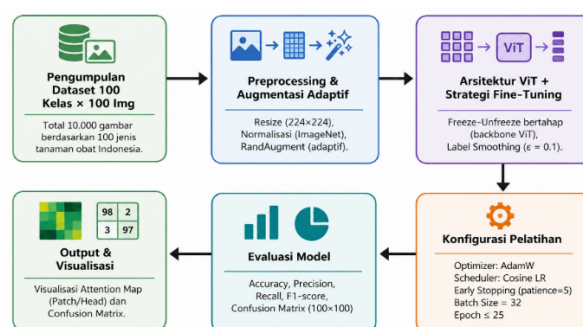
2. Metode Penelitian

2.1. Analisis Masalah

Identifikasi 100 jenis tanaman obat Indonesia secara manual menghadapi tantangan signifikan akibat kemiripan morfologi antar spesies, variasi pencahayaan, latar belakang yang tidak terkontrol, serta keterbatasan ahli botani di lapangan [10], [11]. Secara komputasional, pendekatan konvensional berbasis CNN cenderung mengandalkan *receptive field* lokal sehingga kurang optimal dalam menangkap dependensi spasial jangka panjang antar bagian tanaman. Di sisi lain, ViT menawarkan kemampuan pemodelan konteks global melalui mekanisme *self-attention*, namun arsitektur ini sangat bergantung pada data berskala besar [12], [13]. Pada skenario penelitian ini, *dataset* bersifat menengah (10.000 citra, 100 kelas $\times$ 100 citra/kelas), sehingga penerapan ViT secara langsung berisiko mengalami *overfitting*, ketidakstabilan gradien, dan *catastrophic forgetting* pada bobot pralatih [14]. Oleh karena itu, diperlukan desain *pipeline* yang secara khusus mengintegrasikan strategi *fine-tuning* bertahap, regularisasi terarah, serta augmentasi yang disesuaikan dengan karakteristik morfologi botani agar model dapat belajar pola taksonomi secara stabil dan dapat direproduksi.

2.2. Desain Penelitian

Penelitian ini dirancang melalui alur sistematis yang menghubungkan pengolahan data, arsitektur model, konfigurasi pelatihan, hingga evaluasi performa. Desain penelitian disajikan dalam Gambar 1.



Gambar 1. Desain Penelitian

Tahapan pertama adalah pengumpulan *dataset*. Tahap ini dilakukan dengan menghimpun data citra tanaman obat yang berasal dari repositori publik dan dokumentasi primer. Seluruh data kemudian melalui proses validasi kualitas untuk memastikan bahwa setiap

kelas memiliki representasi visual yang realistis, beragam, dan sesuai dengan karakteristik asli tanaman sehingga dapat mendukung proses pelatihan model secara optimal.

Tahapan yang kedua adalah *preprocessing* dan augmentasi adaptif. Pada tahapan ini, seluruh citra disesuaikan dimensinya dan dinormalisasi agar memiliki format yang seragam sebelum digunakan dalam pelatihan model. Selanjutnya diterapkan augmentasi adaptif untuk meningkatkan variasi data tanpa menghilangkan ciri diagnostik botani penting, seperti pola venasi daun, tekstur, maupun warna bunga, sehingga karakteristik visual tanaman tetap terjaga [15].

Tahapan yang ketiga adalah penyusunan arsitektur ViT dan *fine-tuning*. Model yang digunakan adalah ViT dengan arsitektur `vit_base_patch16_224` yang diinisialisasi menggunakan bobot pra-latih dari ImageNet. Proses pelatihan dilakukan menggunakan strategi *fine-tuning* dengan pembekuan lapisan secara bertahap agar pengetahuan umum yang telah dipelajari model tetap dipertahankan sambil menyesuaikan kemampuan model terhadap domain klasifikasi tanaman obat [16].

Tahapan yang keempat adalah konfigurasi pelatihan. Konfigurasi pelatihan dirancang untuk menangani skenario *dataset* berukuran menengah dengan jumlah kelas yang banyak. Optimasi *hyperparameter* dilakukan untuk memperoleh performa terbaik, didukung dengan penggunaan penjadwalan *learning rate* berbasis kosinus serta mekanisme *early stopping* guna meningkatkan stabilitas pelatihan dan mencegah terjadinya *overfitting* [17].

Tahapan terakhir adalah evaluasi dan output. Tahap evaluasi dilakukan dengan mengukur performa model menggunakan berbagai metrik kuantitatif untuk mengetahui tingkat akurasi dan kemampuan generalisasi model. Selain itu, dilakukan analisis menggunakan *confusion matrix* untuk mengevaluasi pola kesalahan klasifikasi antar kelas serta memastikan bahwa keputusan klasifikasi didasarkan pada karakteristik botani yang relevan.

2.3. Pengumpulan dan Penyiapan Dataset

Dataset terdiri dari 10.000 citra yang mencakup 100 spesies tanaman obat Indonesia. Setiap kelas berisi tepat 100 citra dengan variasi sudut pengambilan, latar belakang, dan intensitas pencahayaan. Pembagian dataset dilakukan menggunakan metode random split pada tingkat keseluruhan dataset (non-stratified split) dengan rasio 70% data latih, 15% data validasi, dan 15% data uji. Dengan total 10.000 citra, proses ini menghasilkan 7.000 citra latih, 1.500 citra validasi, dan 1.500 citra uji. Karena pembagian dilakukan secara acak tanpa mempertahankan proporsi setiap kelas secara terpisah, jumlah citra pada masing-masing kelas dalam subset validasi maupun data uji tidak selalu

sama. Namun demikian, proporsi total dataset tetap dipertahankan sesuai rasio 70:15:15.

Seluruh proses pembagian data dilakukan menggunakan `seed = 42` untuk menjamin reproduktibilitas eksperimen. Pendekatan ini menghasilkan distribusi kelas yang sedikit bervariasi pada data validasi dan data uji sehingga evaluasi dilakukan menggunakan metrik macro precision, macro recall, dan macro F1-score yang tidak dipengaruhi dominasi kelas tertentu. Gambar 2 merupakan contoh citra tanaman obat Indonesia yang digunakan pada penelitian ini.



Gambar 2. *Dataset* Penelitian

2.4. *Preprocessing* dan Augmentasi Adaptif

Pada tahap *preprocessing*, seluruh citra diubah ukurannya menjadi 224×224 piksel agar sesuai dengan dimensi input yang dibutuhkan oleh arsitektur ViT, sehingga setiap citra dapat diproses secara konsisten oleh model. Citra dinormalisasi menggunakan statistik distribusi ImageNet dengan nilai $mean = [0,485; 0,456; 0,406]$ dan $standard\ deviation = [0,229; 0,224; 0,225]$. Penggunaan nilai ini dilakukan karena model `vit_base_patch16_224` telah melalui proses pra-latih pada *dataset* ImageNet, sehingga distribusi data masukan perlu disesuaikan agar sejalan dengan distribusi data saat model pertama kali dilatih. Normalisasi tersebut diterapkan pada seluruh *subset* data (latih, validasi, dan uji) untuk mempercepat konvergensi pelatihan, menjaga stabilitas gradien, serta meningkatkan efektivitas transfer pengetahuan (*transfer learning*) dari model pra-latih ke domain tanaman obat.

Untuk meningkatkan kemampuan generalisasi model tanpa merusak karakteristik morfologi tanaman, diterapkan *pipeline* augmentasi khusus hanya pada data latih yang meliputi *RandomResizedCrop* dengan skala 0,6–1,0 dan rasio aspek 0,8–1,2, *RandomHorizontalFlip* dengan probabilitas 0,5, *ColorJitter* untuk mengatur kecerahan, kontras, dan saturasi hingga $\pm 20\%$, serta pergeseran *hue* sebesar $\pm 0,1$ untuk mensimulasikan variasi pencahayaan alami tanpa mengubah warna diagnostik tanaman secara signifikan, dan *RandAugment* dengan dua operasi dan *magnitude* 9. Augmentasi seperti *VerticalFlip* tidak digunakan karena berpotensi mengubah orientasi daun yang asimetris. Pergeseran *hue* dibatasi pada nilai kecil ($\pm 0,1$) untuk tetap mempertahankan warna diagnostik

yang relevan dalam identifikasi botani. Seluruh rancangan augmentasi ini disesuaikan secara khusus untuk domain tanaman obat agar model lebih fokus mempelajari karakteristik morfologi tanaman dibandingkan artefak latar belakang citra.

2.5. Arsitektur *Vision Transformer* dan Strategi *Fine-tuning*

Model inti yang digunakan dalam penelitian ini adalah *vit base_patch16_224* yang telah melalui proses pelatihan pada *dataset* ImageNet-1k. Arsitektur ini bekerja dengan membagi citra input menjadi 196 *patch* berukuran 16×16 piksel, yang masing-masing diproyeksikan menjadi vektor berdimensi 768. Melalui mekanisme *multi-head self-attention*, model mampu menangkap hubungan spasial global antar-*patch*, sedangkan token [CLS] berperan sebagai *aggregator* fitur akhir yang merepresentasikan keseluruhan citra. Mengingat *dataset* yang digunakan berskala menengah, diterapkan strategi *fine-tuning* bertahap yang dirancang khusus untuk menjamin stabilitas pelatihan dan reproduktibilitas eksperimen. Tahap awal diawali dengan penggantian *classifier head* bawaan menjadi rangkaian *LayerNorm*, *Dropout* (0.3), dan *Linear* (768 → 100) untuk menyesuaikan dimensi output dengan jumlah kelas. Pada tiga *epoch* pertama, seluruh parameter *backbone* dibekukan (*requires_grad=False*) dan hanya head baru yang diperbarui dengan laju pembelajaran 2×10^{-5} , bertujuan menstabilkan gradien sebelum adaptasi skala penuh. Selanjutnya, pada *epoch* ke-4 hingga 25, *backbone* dibuka sepenuhnya untuk proses *full fine-tuning* menggunakan *optimizer* AdamW dengan *weight decay* sebesar 0,05. Sebagai bentuk regularisasi tambahan, fungsi kerugian yang digunakan adalah *CrossEntropyLoss* yang dilengkapi dengan label *smoothing* ($\epsilon=0,1$) guna mencegah kecenderungan *overconfidence* model dan meningkatkan keakuratan kalibrasi probabilitas.

2.6. Konfigurasi Pelatihan dan Evaluasi

Proses pelatihan model dilaksanakan dengan serangkaian konfigurasi *hyperparameter* yang dirancang untuk menjamin stabilitas konvergensi serta mencegah *overfitting*. Sebagai *optimizer*, digunakan AdamW dengan laju pembelajaran *learning rate* sebesar 2×10^{-5} , parameter betas (0,9; 0,999), dan *weight decay* 0,05 yang berfungsi sebagai regularisasi bobot. *Learning rate* diatur secara dinamis menggunakan penjadwal *CosineAnnealingLR* dengan periode maksimum 25 *epoch*. Pelatihan dijalankan menggunakan ukuran *batch* sebesar 32 dan dibatasi maksimal 25 *epoch*, dilengkapi dengan mekanisme *early stopping* yang memiliki *patience* 5 *epoch* berdasarkan pemantauan akurasi validasi, sehingga pelatihan akan berhenti secara otomatis apabila performa model tidak lagi menunjukkan peningkatan. Untuk memastikan hasil yang konsisten dan dapat direproduksi, seluruh proses dijalankan secara deterministik dengan menetapkan *seed* 42 dan

mengaktifkan mode deterministik pada *backend cuDNN PyTorch*. Setelah fase pelatihan selesai, performa model dievaluasi secara komprehensif menggunakan metrik standar klasifikasi citra, yaitu akurasi, presisi, *recall*, dan skor F1 (*macro* dan *weighted*). Matriks konfusi divisualisasikan untuk mengidentifikasi pasangan kelas yang sering tertukar, sementara ekstraksi fitur CLS token diproyeksikan menggunakan t-SNE untuk memvisualisasikan separabilitas antar kelas dalam ruang fitur. Pendekatan ini memungkinkan analisis apakah model telah mempelajari representasi yang bermakna secara taksonomi, sekaligus memvalidasi bahwa keputusan klasifikasi didasarkan pada karakteristik morfologi yang relevan seperti bentuk daun, venasi, atau struktur bunga, bukan pada artefak latar belakang citra. Seluruh rangkaian eksperimen ini dijalankan di lingkungan Google Colab yang didukung oleh GPU NVIDIA T4/A100, sehingga menjamin keterjangkauan sumber daya komputasi sekaligus memastikan reproduktibilitas hasil penelitian secara ilmiah.

3. Hasil dan Pembahasan

3.1. Pengembangan dan Konfigurasi Sistem

Seluruh proses pelatihan model dilaksanakan di lingkungan *Google Colaboratory* dengan akselerator GPU NVIDIA yang mendukung CUDA. Konfigurasi *runtime* yang digunakan meliputi PyTorch versi 2.10.0+cu128 dan TIMM versi 1.0.26. *Dataset* yang digunakan adalah *Indonesia Medicinal Plant Dataset* dalam format ZIP berukuran sekitar 94 MB, yang berhasil diekstraksi secara otomatis oleh *pipeline* pengolahan data. Setelah proses ekstraksi, sistem mendeteksi 100 kelas tanaman obat secara otomatis. Konfigurasi *hyperparameter* yang digunakan pada eksperimen ini adalah *batch size* 32, jumlah *epoch* maksimal 25, *learning rate* 2×10^{-5} , dan rasio data validasi sebesar 0,15. *Dataset* kemudian dibagi menjadi 7.000 citra latih (70%), 1.500 citra validasi (15%), dan 1.500 citra uji (15%) sesuai proporsi 70:15:15 yang telah dijelaskan pada Sub-bab 2.3.

3.2. Arsitektur Model

Model yang dibangun berbasis arsitektur ViT dengan *backbone vit base_patch16_224* yang diinisialisasi menggunakan bobot pelatihan dari ImageNet-1k. Total parameter model adalah 85.877.092, seluruhnya bersifat *trainable* selama fase *full fine-tuning*. Arsitektur *classifier head* yang ditambahkan terdiri dari *LayerNorm*, *Dropout* (0.3), dan lapisan *Linear* berukuran 768→100 untuk menyesuaikan dimensi output dengan jumlah kelas. *Patch embedding* menghasilkan representasi tensor berukuran [1, 196, 768] yang merepresentasikan 196 *patch* berukuran 16×16 piksel dari citra input 224×224. Estimasi total ukuran model mencapai 505,69 MB, dengan ukuran parameter sebesar 342,90 MB dan ukuran aktivasi *forward/backward* sebesar 162,19 MB. Hasil dari

proses pembangunan arsitektur ViT dapat dilihat pada Gambar 3.

Layer (type:depth-idx)	Output Shape	Param #
ViTFeatureExtractor	[1, 100]	--
ViTTransformer: 1-1	--	152,064
PatchEmbed: 2-1	[1, 196, 768]	590,592
Dropout: 2-2	[1, 197, 768]	--
Identity: 2-3	[1, 197, 768]	--
Identity: 2-4	[1, 197, 768]	--
Sequential: 2-5	[1, 197, 768]	85,054,464
LayerNorm: 2-6	[1, 197, 768]	1,536
Sequential: 1-2	[1, 100]	--
LayerNorm: 2-7	[1, 768]	1,536
Dropout: 2-8	[1, 768]	--
Linear: 2-9	[1, 100]	76,900
Total params: 85,877,092		
Trainable params: 85,877,092		
Non-trainable params: 0		
Total mult-adds (Units:MEGABYTES): 200.89		
Input size (MB): 0.60		
Forward/backward pass size (MB): 162.19		
Params size (MB): 342.90		
Estimated Total Size (MB): 505.69		

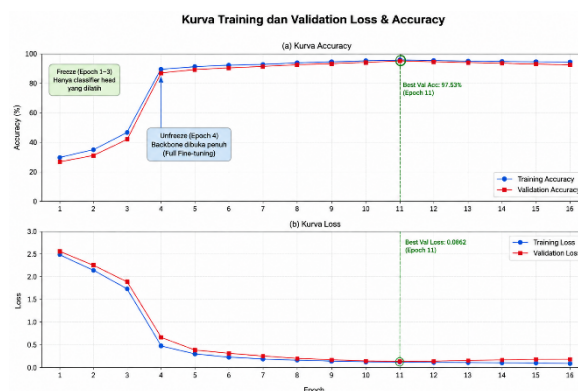
Gambar 3. Arsitektur *Vision Transformer*

Gambar 3 menyajikan ringkasan struktur dan statistik arsitektur model ViT yang dibangun. Bagian atas gambar menampilkan hierarki lapisan model secara lengkap. *Backbone* utama berupa *Vision Transformer* dengan kedalaman 1-1, di dalamnya terdapat modul *PatchEmbed* (2-1) yang bertugas memecah citra menjadi patch berukuran 16×16 piksel dengan output shape [1, 196, 768] dan memiliki 590.592 parameter. Di bawahnya terdapat lapisan *Dropout* (2-2), diikuti oleh tiga lapisan *Identity* (2-3, 2-4, 2-5) yang masing-masing menghasilkan *output shape* [1, 197, 768]. Selanjutnya, lapisan *LayerNorm* (2-6) dengan 1.536 parameter memproses representasi tersebut sebelum diteruskan ke blok *Sequential* (1-2) yang merupakan *classifier head* yang ditambahkan, terdiri atas *LayerNorm* (2-7) dengan 1.536 parameter dan lapisan *Linear* (2-9) berukuran 768→100 dengan 76.900 parameter. Bagian bawah gambar menampilkan statistik keseluruhan model: total parameter sebesar 85.877.092, seluruhnya bersifat trainable dengan 0 parameter yang dibekukan. Ukuran input model sebesar 0,60 MB, ukuran parameter sebesar 342,90 MB, dan ukuran aktivasi *forward/backward* sebesar 162,19 MB, sehingga estimasi total ukuran model mencapai 505,69 MB.

3.3. Proses Pelatihan

Proses pelatihan model berlangsung selama 16 *epoch* sebelum dihentikan secara otomatis oleh mekanisme *early stopping* dengan *patience* 5, meskipun batas maksimal pelatihan ditetapkan sebesar 25 *epoch*. *Epoch* merupakan satu siklus penuh di mana model mempelajari seluruh 7.000 citra latih, sehingga selama 16 *epoch* model telah menjalani 16 putaran pembelajaran. Mekanisme *early stopping* berfungsi untuk menghentikan pelatihan lebih awal apabila model tidak lagi menunjukkan peningkatan akurasi validasi, guna mencegah terjadinya *overfitting* yaitu kondisi di mana model terlalu menghafalkan data latih sehingga kemampuan generalisasinya terhadap data baru menurun. Dengan nilai *patience* 5, model diberikan toleransi sebanyak 5 *epoch* berturut-turut tanpa peningkatan sebelum pelatihan dihentikan. Dalam eksperimen ini, akurasi validasi terbaik tercapai

pada *epoch* ke-11 sebesar 97,53%, dan karena tidak ada peningkatan lebih lanjut hingga *epoch* ke-16, pelatihan dihentikan secara otomatis dengan model terbaik yang telah tersimpan pada *epoch* ke-11. Hasil dari proses ini dapat dilihat pada Gambar 4.



Gambar 4. Hasil *Training*

Gambar 4 menampilkan kurva akurasi dan loss selama proses pelatihan 16 *epoch*. Pada fase pembekuan (*epoch* 1–3), akurasi validasi masih rendah sekitar 28,00% karena hanya *classifier head* yang diperbarui. Lonjakan signifikan terjadi pada *epoch* ke-4 saat backbone dibuka penuh, dengan akurasi validasi melonjak menjadi 91,53%. Kurva terus meningkat hingga mencapai puncaknya pada *epoch* ke-11 sebesar 97,53%, setelah itu mekanisme *early stopping* menghentikan pelatihan pada *epoch* ke-16 karena tidak ada peningkatan lebih lanjut. Kurva *training loss* yang menurun secara stabil mengonfirmasi konvergensi berlangsung tanpa gejala *overfitting* yang signifikan.

3.4. Evaluasi pada Data Uji

Model terbaik yang tersimpan pada *epoch* ke-11 dengan akurasi validasi sebesar 97,53% selanjutnya dievaluasi menggunakan 1.500 data uji yang telah dipisahkan sejak awal dan tidak pernah digunakan selama proses pelatihan maupun validasi. Pemilihan bobot model pada *epoch* ke-11 dilakukan karena mekanisme *checkpoint* secara otomatis menyimpan bobot terbaik setiap kali akurasi validasi mencapai nilai tertinggi, sehingga bobot yang digunakan untuk evaluasi akhir bukan berasal dari *epoch* terakhir melainkan dari *epoch* dengan performa validasi optimal. Penggunaan data uji yang sepenuhnya terpisah ini bertujuan untuk mengukur kemampuan generalisasi model secara objektif terhadap data yang belum pernah dilihat sebelumnya, sehingga hasil evaluasi benar-benar mencerminkan performa model di kondisi nyata. Hasil evaluasi secara keseluruhan mencakup empat metrik utama yaitu *accuracy*, *precision*, *recall*, dan *F1-score* yang selengkapnya disajikan pada Tabel 1.

Tabel 1. Hasil Evaluasi Data Uji

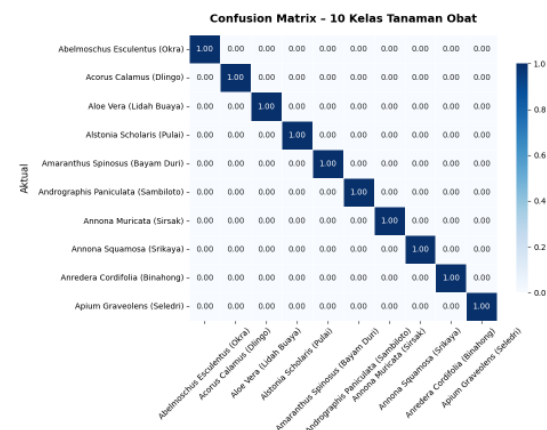
Metrik	Macro Avg	Weighted Avg
<i>Precision</i>	0,974	0,975
<i>Recall</i>	0,975	0,973
<i>F1-score</i>	0,973	0,973
<i>Accuracy</i>	0,973	—

Model mencapai akurasi keseluruhan sebesar 97,3% pada data uji, dengan *macro F1-score* sebesar 0,973 dan *weighted F1-score* sebesar 0,973. Nilai *macro precision* sebesar 0,974 dan *macro recall* sebesar 0,975 menunjukkan bahwa model mampu mengenali hampir seluruh kelas tanaman obat dengan keseimbangan yang baik antara presisi dan sensitivitas, meskipun distribusi jumlah citra per kelas pada data uji tidak sepenuhnya seragam. Ketidakseragaman tersebut terjadi karena pembagian dataset dilakukan menggunakan fungsi *random_split* pada tingkat keseluruhan dataset (*non-stratified random split*), sehingga pembagian hanya mempertahankan proporsi total data sebesar 70:15:15 tanpa menjamin jumlah sampel yang sama pada setiap kelas. Akibatnya, jumlah citra pada setiap kelas di subset pengujian bervariasi antara 7–24 citra, sedangkan jumlah keseluruhan data uji tetap sebanyak 1.500 citra. Hasil tersebut menunjukkan bahwa kombinasi *Vision Transformer*, *strategi stepwise fine-tuning* berbasis *freeze-unfreeze*, *label smoothing*, *cosine learning rate scheduling*, dan augmentasi adaptif mampu menghasilkan performa klasifikasi yang tinggi pada dataset tanaman obat Indonesia. Namun, penelitian ini belum melakukan *ablation study* maupun perbandingan langsung dengan *strategi fine-tuning* standar atau arsitektur CNN lainnya, sehingga kontribusi masing-masing komponen, termasuk strategi *freeze-unfreeze*, belum dapat dievaluasi secara terpisah. Oleh karena itu, penelitian selanjutnya perlu melakukan studi komparatif dan *ablation study* untuk mengukur kontribusi setiap komponen terhadap peningkatan kinerja model.

Secara per-kelas, sebagian besar dari 100 spesies tanaman obat berhasil diklasifikasikan dengan sempurna (*F1-score* = 1,000), diantaranya *Abelmoschus Esculentus* (Okra), *Annona Muricata* (Sirsak), *Alstonia Scholaris* (Pulai), *Curcuma Domestica* (Kunyit), dan *Clitoria Ternatea* (Kembang Telang). Performa tertinggi pada kelas dengan ciri morfologi yang sangat khas dan berbeda dari spesies lain.

Beberapa kelas yang menunjukkan performa lebih rendah antara lain *Clinalanthus Nutans* (Dandang Gendis) dengan *F1-score* 0,833, *Dioscorea Hispida* (Gadung) dengan *F1-score* 0,882, dan *Melissa officinalis* (Lemon Balm) dengan *F1-score* 0,889. Rendahnya performa pada kelas-kelas ini kemungkinan disebabkan oleh kemiripan morfologi dengan spesies lain, variasi pencahayaan yang lebih tinggi pada citra, atau jumlah data uji yang lebih sedikit untuk kelas tersebut. Sebagai contoh, *Melissa officinalis* memiliki bentuk daun yang secara visual mirip dengan beberapa spesies herbal lainnya sehingga lebih rentan terhadap kesalahan klasifikasi. Visualisasi hasil prediksi model terhadap seluruh 100 kelas pada data uji ditampilkan

dalam bentuk *confusion matrix* pada Gambar 5.



Gambar 5. Confusion matrix

Berdasarkan Gambar 5 secara keseluruhan, matriks menunjukkan dominasi nilai tinggi pada diagonal utama, mengindikasikan bahwa sebagian besar kelas berhasil diklasifikasikan dengan benar. Kesalahan klasifikasi yang tampak umumnya terjadi antara spesies dengan kemiripan morfologi tinggi, seperti *Clinalanthus Nutans* dan *Melissa officinalis*.

3.5. Uji Prediksi Interaktif

Untuk memvalidasi kemampuan generalisasi model secara langsung, dilakukan uji prediksi pada citra tanaman jahe (tanaman-jahe.jpg) yang tidak termasuk dalam *dataset* pelatihan maupun pengujian. Model berhasil mengidentifikasi gambar tersebut sebagai *Zingiber Officinale* (Jahe) sebagai prediksi pertama dengan *confidence* sebesar 74,99%. Prediksi kedua dan ketiga adalah *Curcuma Domestica* (Kunyit) dengan *confidence* 2,02% dan *Hibiscus Rosa-sinensis* (Kembang Sepatu) dengan *confidence* 0,67%.

Selisih *confidence* yang sangat besar antara prediksi pertama (74,99%) dan prediksi kedua (2,02%) menunjukkan bahwa model memiliki tingkat keyakinan yang tinggi terhadap identifikasi jahe. Hal ini juga relevan secara botani, karena jahe dan kunyit memang termasuk dalam famili yang sama (*Zingiberaceae*) dengan karakteristik morfologi daun yang serupa, sehingga wajar apabila model menempatkan kunyit sebagai kandidat alternatif terdekat.

4. Kesimpulan

Penelitian ini telah berhasil membangun sistem identifikasi 100 jenis tanaman obat Indonesia menggunakan arsitektur ViT dengan pendekatan *fine-tuning* bertahap dan augmentasi citra adaptif. Berdasarkan hasil eksperimen yang telah dilakukan, dapat ditarik beberapa kesimpulan sebagai berikut.

Pertama, Penelitian ini menunjukkan bahwa integrasi *Vision Transformer* dengan *stepwise fine-tuning*, *label smoothing*, *cosine learning rate scheduling*, dan augmentasi adaptif mampu menghasilkan performa

klasifikasi yang tinggi pada dataset tanaman obat Indonesia.

Kedua, model yang dibangun mencapai akurasi sebesar 97,3% pada 1.500 data uji dengan *macro F1-score* 0,973, *macro precision* 0,974, dan *macro recall* 0,975. Hasil ini menunjukkan bahwa model memiliki kemampuan generalisasi yang sangat baik dan mampu mengidentifikasi hampir seluruh spesies tanaman obat dengan tingkat kepercayaan yang tinggi, sebagaimana dikonfirmasi melalui uji prediksi interaktif pada citra Jahe dengan *confidence* 74,99%.

Ketiga, *pipeline* augmentasi adaptif yang dirancang khusus untuk domain botani, dengan membatasi pergeseran *hue* pada nilai kecil ($\pm 0,1$) dan tidak menggunakan *VerticalFlip*, terbukti mampu meningkatkan kemampuan generalisasi model tanpa merusak ciri morfologi diagnostik tanaman yang penting untuk proses identifikasi.

Secara keseluruhan, pendekatan yang diusulkan dalam penelitian ini memberikan solusi yang efektif, stabil, dan dapat direproduksi untuk permasalahan klasifikasi tanaman obat berskala banyak kelas pada dataset berukuran menengah. Penelitian ini menggunakan metode *random split (non-stratified)* sehingga distribusi jumlah sampel pada setiap kelas di data validasi dan data uji tidak sepenuhnya seragam. Pada penelitian selanjutnya, penggunaan *stratified split* dapat dipertimbangkan agar jumlah sampel setiap kelas lebih seimbang, sehingga evaluasi performa per kelas menjadi lebih konsisten.

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan dalam menginterpretasikan hasil serta menjadi arah pengembangan selanjutnya. Pertama, evaluasi pada penelitian ini belum menyertakan perbandingan langsung terhadap arsitektur baseline lain seperti *ResNet50*, *EfficientNet*, atau ViT standar tanpa strategi *fine-tuning* bertahap, sehingga klaim keunggulan pendekatan yang diusulkan bersifat internal (dibandingkan terhadap variasi konfigurasi pelatihan ViT itu sendiri) dan belum divalidasi secara komparatif lintas arsitektur. Studi lanjutan yang menyertakan tabel perbandingan kuantitatif terhadap beberapa arsitektur CNN dan ViT standar pada dataset yang identik akan memperkuat validitas klaim performa. Kedua, penelitian ini belum menyertakan analisis interpretabilitas berbasis visualisasi attention map untuk mengonfirmasi bahwa keputusan klasifikasi model didasarkan pada area morfologi yang relevan secara botani (misalnya bentuk daun atau pola venasi) dan bukan pada artefak latar belakang; interpretabilitas pada penelitian ini baru dievaluasi secara tidak langsung melalui pola kesalahan pada confusion matrix. Visualisasi attention map menjadi arah pengembangan yang direkomendasikan untuk penelitian selanjutnya. Ketiga, Penelitian ini menggunakan non-stratified random split, sehingga jumlah sampel pada setiap kelas di data validasi maupun data uji tidak sepenuhnya

seragam. Walaupun kondisi tersebut tidak memengaruhi proporsi total dataset, penerapan stratified split pada penelitian selanjutnya akan menghasilkan distribusi sampel yang lebih seimbang pada setiap kelas sehingga evaluasi performa per kelas dapat dilakukan dengan lebih akurat. Keempat, uji prediksi interaktif pada penelitian ini baru diujikan pada satu citra di luar dataset (citra Jahe), sehingga generalisasi model terhadap citra dengan kondisi lapangan yang lebih beragam (pencahayaan ekstrem, latar belakang kompleks, sudut pengambilan tidak ideal) masih perlu diuji lebih lanjut pada set data eksternal yang lebih besar.

Daftar Rujukan

- [1] D. R. P. Mahendri and T. Y. Hadiwandura, "Implementasi Deep Learning Untuk Identifikasi Tanaman Rimpang Menggunakan Metode Convolutional Neural Network," *J. Comput. Sci. Inf. Technol.*, vol. 6, no. 1, pp. 1–8, 2025, doi: 10.37859/coscitech.v6i1.8943.
- [2] M. Alif Fathan, T. Sugiharto, and I. Lesmana, "Klasifikasi jenis tanaman philodendron berdasarkan citra daun menggunakan algoritma CNN," *J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 140–147, 2025, doi: 10.37859/coscitech.v6i2.9484.
- [3] N. Trisanti, N. T. Romadloni, and N. H. Sya'bani, "Klasifikasi Citra Buah Menggunakan Algoritma K-Nearest Neighbour (KNN) dan Metode Euclidean Distance," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 3, pp. 8306–8312, 2025, doi: 10.31004/riggs.v4i3.3243.
- [4] R. Pujiati and N. Rochmawati, "Identifikasi Citra Daun Tanaman Herbal Menggunakan Metode Convolutional Neural Network (CNN)," *J. Informatics Comput. Sci.*, vol. 3, no. 03, pp. 351–357, 2022, doi: 10.26740/jinacs.v3n03.p351-357.
- [5] B. Setiyono *et al.*, "Identifikasi Tanaman Obat Indonesia Melalui Citra Daun Menggunakan Metode Convolutional Neural Network (CNN)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 2, pp. 385–392, 2023, doi: 10.25126/jtiik.20236809.
- [6] S. A. E. Albakia and R. A. Saputra, "IDENTIFIKASI JENIS DAUN TANAMAN OBAT MENGGUNAKAN METODE CONVOLUTIONAL NEURAL NETWORK (CNN) DENGAN MODEL VGG16," *J. Inform. Polinema*, vol. 9, no. 4, pp. 451–460, 2023, doi: 10.33795/jip.v9i4.1420.
- [7] A. Dosovitskiy *et al.*, "an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021, doi: <https://doi.org/10.48550/arXiv.2010.11929>.
- [8] C. F. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 347–356, 2021, doi: 10.1109/ICCV48922.2021.00041.
- [9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.
- [10] M. S. Ikrar Musyaffa, N. Yudistira, M. A. Rahman, A. H. Basori, A. B. Firdausiah Mansur, and J. Batoro, "IndoHerb: Indonesia medicinal plants recognition using transfer learning and deep learning," *Heliyon*, vol. 10, no. 23, pp. 1–34, 2024, doi: 10.1016/j.heliyon.2024.e40606.
- [11] P. S. Thakur, P. Khanna, T. Sheorey, and A. Ojha, "Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT," 2022, [Online]. Available: <http://arxiv.org/abs/2207.07919>
- [12] A. Rachman, A. Fauzi, and B. Wijonarko, "Klasifikasi Citra Rempah-Rempah Di Indonesia Menggunakan Vision Transformer," *JTIK (Jurnal Tek. Inform. Kaputama)*, vol. 10, no. 1, pp. 16–27, 2026, doi: 10.59697/jtik.v10i1.1168.

-
- [13] Darmatasia and M. H. H, "Model Efisien Berbasis Mobile Vision Transformer (Mobilevit) Untuk Klasifikasi Jenis Tanah Dari Citra," *J. INSTEK (Informatika Sains dan Teknol.*, vol. 10, no. 2, pp. 518–533, 2025, doi: 10.24252/instek.v10i2.61891.
- [14] A. Hasbi and B. Tumanggor, "Implementation of the Vision Transformer (ViT) Method in Soil Type Classification," *J. Ilm. Sist. Inf.*, vol. 5, no. 1, pp. 574–583, 2026, doi: 10.1109/SCES55490.2022.9887636.
- [15] R. Uthama, B. Hendrik, M. T. Informatika, F. I. Komputer, U. P. Indonesia, and U. M. Riau, "Vision Transformer untuk Identifikasi 15 Variasi Citra Ikan Koi Rayhan," *J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 159–168, 2024, doi: 10.37859/coscitech.v5i1.6711.
- [16] R. R. Ar, Agusriyati, and S. Moka, "Pendeteksian Dini Stunting Pada Balita Menggunakan Vision Transformer (VIT) Berbasis Citra Tubuh," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3S1, pp. 896–902, 2025, doi: 10.23960/jitet.v13i3S1.7888.
- [17] H. Wu *et al.*, "CvT: Introducing Convolutions to Vision Transformers," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 22–31, 2021, doi: 10.1109/ICCV48922.2021.00009.