JITKOM

Jurnal Ilmu Teknik dan Komputer
Vol. 10. No. 1, Januari 2026: 20-25
p-ISSN: 2548-740X e-ISSN: 2621-1491

Klasifikasi Citra untuk Menentukan Daun Segar atau Daun Layu Menggunakan *Vision Transformer* (ViT) untuk Otomatisasi Proses Penyortiran Daun

Muhammad Aryaka Zamzami*, Rivaldiansyah Pramadhan, Fariez Harman

Teknik Informatika, Universitas Mercu Buana, Jakarta
*m.aryakazamzam@gmail.com

***Abstrak*— Penelitian ini mengeksplorasi pemanfaatan arsitektur Vision Transformer (ViT) untuk klasifikasi citra daun ke dalam dua kategori, yaitu daun segar dan daun layu. Data yang digunakan berjumlah 3000 citra (gambar), diperoleh secara langsung melalui kamera smartphone, masing-masing terdiri dari 1.500 gambar daun segar dan 1.500 gambar daun layu. Model ViTForImageClassification dari Hugging Face Transformers dipilih sebagai kerangka utama, diimplementasikan menggunakan PyTorch dalam platform Google Colaboratory. 10% dataset digunakan sebagai data test guna menilai performa model yang telah melalui proses pelatihan sebelumnya sebesar 70% data train dan 20% data validasi. Berdasarkan hasil evaluasi, model ViT mampu melakukan klasifikasi dengan akurasi keseluruhan sebesar 88,9%, precision 90%, recall 89%, serta F1-score sebesar 89%. Temuan ini mengindikasikan bahwa pendekatan berbasis deep learning, khususnya Vision Transformer, memiliki potensi signifikan dalam mendukung pemantauan kesehatan tanaman dan pengendalian mutu produk pertanian.**

Kata Kunci*— *Daun Segar; Daun Layu; Klasifikasi Citra; Pengolahan Citra Digital; Vision Transformer

DOI: 10.22441/jitkom.v10i1.004

Article History:
Received: July 27, 2025
Revised: Oct 30, 2025
Accepted: Nov 25, 2025
Published: Jan 31, 2026

I. PENDAHULUAN

Kemampuan untuk membedakan antara daun yang masih segar dan yang sudah layu memiliki peran penting dalam berbagai penerapan praktis, khususnya di bidang pertanian modern [2]. Deteksi kondisi daun secara tepat dapat membantu pemantauan kesehatan tanaman secara langsung, mendukung pengambilan keputusan terkait irigasi dan pemupukan, serta berkontribusi dalam menjaga kualitas hasil panen. Dalam kajian botani, teknologi klasifikasi otomatis semacam ini juga berpotensi mempercepat proses penelitian dan pengawasan terhadap ekosistem tumbuhan [6].

Selama ini, identifikasi kondisi daun umumnya masih mengandalkan pengamatan visual secara manual oleh ahli botani atau petani yang berpengalaman. Namun, pendekatan tersebut memiliki berbagai keterbatasan, seperti penilaian yang subjektif, ketergantungan pada tenaga ahli, serta hasil yang bisa kurang konsisten, terutama jika harus menangani data dalam jumlah besar. Tantangan-tantangan inilah yang mendorong lahirnya sistem otomatis yang mampu memberikan hasil klasifikasi yang lebih objektif dan stabil.

Kemajuan teknologi pengolahan citra digital berbasis deep learning telah membuka peluang besar dalam tugas-tugas klasifikasi citra, termasuk dalam bidang pertanian [3]. Salah satu inovasi terkini adalah arsitektur Vision Transformer (ViT), yang mengadaptasi mekanisme attention dari dunia pemrosesan bahasa alami (NLP) ke dalam bidang visi komputer. ViT telah menunjukkan performa unggul dalam berbagai benchmark klasifikasi gambar [1]. Tidak seperti Convolutional Neural Network (CNN) yang bekerja dengan filter konvolusi secara bertingkat [12], ViT memecah gambar menjadi bagian-bagian kecil (patch) dan memprosesnya sebagai urutan data dengan menggunakan mekanisme self-attention [7].

Penelitian ini berupaya mengembangkan sistem klasifikasi citra otomatis untuk membedakan antara daun segar dan daun layu secara akurat, dengan memanfaatkan arsitektur Vision Transformer. Tujuan utamanya adalah untuk mengimplementasikan dan mengevaluasi performa ViT dalam klasifikasi biner citra daun, dengan harapan temuan ini dapat mendukung pengembangan teknologi pertanian presisi dan sistem monitoring tanaman secara otomatis.

II. LITERATURE REVIEW

Kemajuan pesat dalam bidang *computer vision* telah membuka berbagai peluang aplikasi pengolahan citra dalam sektor pertanian [3], [6]. Salah satu tantangan yang relevan adalah pengklasifikasian kondisi daun tanaman, misalnya membedakan antara daun segar dan layu yang sebelumnya banyak bergantung pada observasi visual manual oleh petani atau ahli botani. Pendekatan tradisional ini memiliki keterbatasan seperti ketergantungan pada tenaga ahli, subjektivitas dalam penilaian, serta potensi ketidakkonsistenan ketika diterapkan dalam skala besar.

Dalam konteks ini, teknologi deep learning muncul sebagai solusi potensial untuk mengotomatisasi proses klasifikasi citra tanaman secara lebih cepat, akurat, dan konsisten. Berbagai studi telah menyoroti penerapan *Convolutional Neural Networks* (CNN) untuk pengenalan penyakit daun atau klasifikasi spesies tanaman. Namun, CNN memiliki keterbatasan dalam menangkap hubungan spasial jangka panjang dan cenderung fokus pada fitur lokal [9].

Sebagai alternatif, Vision Transformer (ViT) diperkenalkan sebagai pendekatan baru yang mentransformasikan gambar menjadi *patch-patch* kecil, yang kemudian diproses sebagai urutan, mirip seperti pemrosesan kata dalam NLP [1]. ViT telah menunjukkan performa superior dalam berbagai benchmark klasifikasi gambar skala besar [1].

Studi oleh Nugroho & Witarto membuktikan bahwa Vision Transformer dapat digunakan untuk klasifikasi daun tanaman obat Indonesia, dan menghasilkan akurasi tinggi dengan pelatihan terbatas [5]. Temuan serupa juga diungkapkan oleh Sharma & Vishwakarma, yang menggunakan ViT untuk mendeteksi defisiensi nutrisi pada daun pisang secara otomatis dan efisien dalam konteks klasifikasi multi-kelas [7].

Dalam hal augmentasi data, menunjukkan bahwa teknik augmentasi seperti rotasi, flipping, perubahan warna, dan distorsi geometris dapat secara signifikan meningkatkan kemampuan generalisasi model deep learning, terutama ketika dataset memiliki variasi visual tinggi namun jumlah terbatas [8].

Dengan demikian, literatur terkini menunjukkan bahwa Vision Transformer memiliki potensi besar untuk menggantikan atau melengkapi pendekatan CNN dalam klasifikasi citra berbasis tanaman, khususnya dalam skenario dengan kompleksitas visual tinggi. Penelitian ini hadir untuk melanjutkan eksplorasi tersebut, dengan fokus pada klasifikasi daun segar atau daun layu sebagai kontribusi konkret bagi pengembangan sistem monitoring tanaman otomatis.

III. METODOLOGI PENELITIAN

A. Penelitian Kuantitatif Eksperimental

Penelitian ini dilaksanakan secara daring (online) menggunakan platform Google Colaboratory (Google Colab) sebagai lingkungan komputasi berbasis cloud. Pemanfaatan Google Colab memungkinkan akses terhadap GPU (Graphics Processing Unit) yang diperlukan untuk pelatihan model *deep learning*.

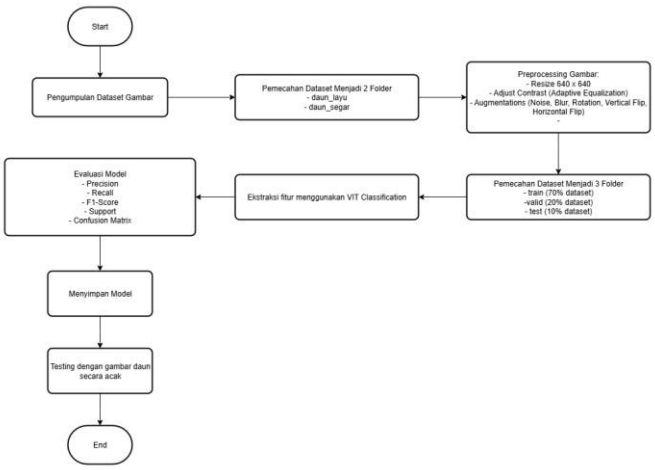
Waktu pelaksanaan penelitian dimulai pada bulan April hingga Juni 2025, yang mencakup tahap pengumpulan data, preprocessing, pelatihan model, hingga evaluasi dan analisis hasil.

Sampel dalam penelitian ini mencakup seluruh citra daun yang dapat merepresentasikan dua kondisi utama: daun segar dan daun layu, yang ditemukan dalam lingkungan sekitar atau kebun sekitar. Sampel diambil secara purposive, yaitu dengan mengumpulkan 3.000 citra daun yang terdiri dari 1.500 citra daun segar dan 1.500 citra daun layu.

Seluruh data sampel dikumpulkan secara langsung oleh peneliti, menggunakan kamera smartphone dalam berbagai kondisi cahaya dan latar belakang, guna mencerminkan variasi alami yang umum ditemukan di lapangan.

B. Akuisisi dan Persiapan Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari citra daun yang diklasifikasikan ke dalam dua kategori, yaitu *daun_segar* dan *daun_layu*. Seluruh data dikumpulkan langsung oleh kami untuk menjamin kualitas dan kesesuaiannya dengan tujuan studi. Proses pengambilan gambar dilakukan menggunakan kamera smartphone masing-masing anggota tim, dengan sengaja dilakukan dalam beragam kondisi pencahayaan dan latar belakang guna meningkatkan variasi dalam dataset.



Gambar 1. Tahapan Penelitian

Pengelompokan gambar dilakukan secara manual, dengan memisahkan citra ke dalam dua folder: “daun_segar” untuk daun yang masih tampak sehat, dan “daun_layu” untuk daun yang menunjukkan gejala kekeringan. Karena sebagian besar gambar awalnya berformat HEIC (High Efficiency Image Container), semua file dikonversi ke format JPG agar kompatibel dengan framework yang digunakan dalam penelitian ini.

Secara keseluruhan, terkumpul sebanyak 3.000 gambar, terdiri dari 1.500 citra daun segar dan 1.500 citra daun layu. Distribusi yang seimbang ini dimaksudkan untuk mencegah bias klasifikasi dalam pelatihan model. Untuk proses evaluasi model, dataset dibagi dengan komposisi 70% untuk pelatihan, 20% untuk validasi, dan 10% untuk pengujian. Dengan skema ini, model ViT dilatih sepenuhnya menggunakan data yang dikumpulkan sendiri, lalu diuji secara menyeluruh pada dataset kustom yang dikembangkan dalam penelitian ini.

Gambar 2. Daun Segar dan daun layu

C. Pra-pemrosesan Data

Dalam konteks deep learning berbasis *computer vision*, data citra tidak dapat langsung diberikan ke model tanpa melalui sejumlah penyesuaian, karena setiap arsitektur memiliki konfigurasi input yang berbeda, baik dari sisi resolusi, format, maupun statistik distribusi piksel [11]. Sebelum citra dapat digunakan sebagai input ke dalam arsitektur Vision Transformer (ViT), diperlukan proses pra-pemrosesan agar format dan karakteristik citra sesuai dengan spesifikasi model.

Tahapan ini merupakan salah satu langkah penting yang menentukan keberhasilan proses pelatihan dan evaluasi model dalam tugas klasifikasi citra.

Salah satu langkah awal dalam pra-pemrosesan adalah penyesuaian ukuran citra. Seluruh gambar diubah ukurannya menjadi 640 x 640 piksel, ukuran yang dianggap optimal untuk sebagian besar model ViT karena mempertahankan detail visual sekaligus sesuai dengan input default arsitektur [4]. Untuk menangani proses ini secara efisien, digunakan komponen "ViTImageProcessor" dari *library* Hugging Face Transformers. Komponen ini dirancang khusus untuk menyelaraskan data masukan dengan kebutuhan model ViT yang telah dilatih sebelumnya (*pre-trained*), termasuk normalisasi statistik piksel, format tensor, serta transformasi lainnya.

Selain pra-pemrosesan standar, teknik augmentasi data juga diterapkan, terutama pada tahap pelatihan, untuk memperkaya keragaman dataset. Augmentasi bertujuan meningkatkan kemampuan generalisasi model terhadap variasi citra di dunia nyata [8]. Dalam implementasinya, Kami menerapkan sejumlah transformasi acak seperti rotasi, flipping horizontal dan vertikal, perubahan warna, translasi, skala, serta distorsi perspektif. Transformasi ini membantu model belajar dari berbagai variasi bentuk dan kondisi pencahayaan daun tanpa perlu menambah jumlah citra secara eksplisit [10].

Distribusi data yang seimbang serta adanya augmentasi pada data pelatihan menjadi salah satu keunggulan pendekatan ini, karena membantu meminimalkan overfitting dan memastikan model dapat menangani variasi yang mungkin ditemui pada data nyata di lapangan. Sementara itu, subset validasi dan pengujian tetap dibiarkan tanpa augmentasi untuk memberikan hasil evaluasi yang objektif terhadap kemampuan generalisasi model.

D. Model Vision Transformer (ViT)

Vision Transformer (ViT) merupakan salah satu inovasi terkini dalam bidang computer vision yang mengadopsi struktur arsitektural dari Transformer, sebuah pendekatan yang awalnya dirancang untuk tugas pemrosesan bahasa alami (natural language processing atau NLP). ViT memperlihatkan bahwa meskipun tidak menggunakan lapisan konvolusional seperti halnya Convolutional Neural Networks (CNN), Transformer tetap mampu menangani tugas klasifikasi citra secara efektif [5]. Bahkan, dalam beberapa skenario, performanya terbukti melampaui CNN pada dataset berskala besar.

Prinsip Kerja dan Struktur ViT, berbeda dengan CNN yang mengekstrak fitur lokal secara bertahap [9], ViT melihat citra sebagai sekumpulan patch yang merepresentasikan bagian-bagian kecil dari keseluruhan gambar, konsep ini mirip dengan bagaimana kalimat direpresentasikan sebagai urutan kata dalam NLP (*Natural Language Processing*). Misalnya, gambar berukuran 224×224 piksel dapat dipecah menjadi patch 16×16 piksel, yang kemudian di-flatten dan diproyeksikan ke dalam ruang vektor berdimensi tetap (embedding space). Agar posisi spasial tetap terjaga, setiap patch diberikan positional encoding sebelum diproses lebih lanjut melalui blok encoder Transformer. Komponen utama dalam arsitektur ViT antara lain:

- *Linear Projection of Image Patches*
- *Positional Embedding*
- *Transformer Encoder Layers (dengan Multi-Head Self-Attention* dan *Feed Forward Network)*
- *Classification Head (Multilayer Perceptron)*

Karena tidak memiliki inductive bias spasial seperti CNN, ViT mengandalkan ukuran dataset yang besar serta teknik regularisasi dan augmentasi yang memadai agar dapat belajar secara efektif.

Penelitian ini mengimplementasikan model ViT menggunakan *ViTForImageClassification* dari *library* Hugging Face Transformers dengan model dasar google/vit-base-patch16-224. Model ini memiliki ukuran patch 16×16 piksel dan menerima citra input berukuran 224×224 piksel. Untuk menyesuaikan dengan kebutuhan klasifikasi biner (daun segar dengan daun layu), bagian akhir (*classification head*) dari model disesuaikan dengan menambahkan lapisan dropout dan *linear layer* baru.

Modifikasi dilakukan pada bagian classifier untuk menambahkan regularisasi dan mencegah *overfitting*, sebagai berikut :

```
self.model.classifier = nn.Sequential(
    nn.Dropout(0.3),
    nn.Linear(self.model.classifier.in_features, num_classes)
)
```

Gambar 3. Modifikasi Classifier

E. Proses Pelatihan Model

Seluruh proses pelatihan dan pengujian model dilakukan menggunakan lingkungan cloud-based Google Colaboratory (Google Colab), yang menyediakan akses ke akselerasi GPU seperti NVIDIA T4 GPU untuk efisiensi komputasi. Lingkungan cloud ini memungkinkan eksperimen dengan model deep learning yang membutuhkan sumber daya komputasi tinggi tanpa memerlukan kepentingan hardware lokal.

Parameter pelatihan (*training*) utama yang digunakan dalam proses fine-tuning atau evaluasi model meliputi

- *Transfer Learning* dan *Fine-Tuning* Bertahap *(Gradual Unfreezing)*

Model ViT yang digunakan merupakan model pre- trained, yaitu telah dilatih sebelumnya pada dataset berskala besar seperti ImageNet-21k. Untuk mengadaptasi model ini pada klasifikasi daun segar dengan daun layu, pendekatan transfer learning diterapkan.

Pada tahap awal pelatihan, seluruh parameter dari backbone ViT dibekukan (*frozen*) untuk mencegah overfitting dan memfokuskan pelatihan pada bagian klasifikasi akhir (*classifier head*). Setelah lima epoch pertama, model backbone dibuka kembali (*unfreeze*) secara bertahap, memungkinkan fine-tuning terhadap seluruh parameter ViT dengan learning rate yang lebih kecil. Langkah ini penting untuk menjaga kestabilan pembelajaran awal dan memanfaatkan pengetahuan umum dari pelatihan sebelumnya secara efektif

- *Label Smoothing*

Untuk mencegah model menjadi terlalu yakin (*overconfident*) terhadap prediksinya, digunakan teknik label *smoothing*. Label *Smoothing* merupakan metode regularisasi di mana label target tidak diberikan secara keras (hard target) berupa nilai 0 atau 1, melainkan diberikan sebagai distribusi probabilistik yang sedikit dilembutkan. Hal ini mendorong model agar tidak mengabaikan kelas lain sepenuhnya, sehingga dapat mengurangi risiko *overfitting* dan meningkatkan kemampuan generalisasi

- *Gradient Clipping*

Untuk mengatasi potensi masalah *exploding gradients*, terutama saat fine-tuning layer Transformer, digunakan teknik gradient clipping. Seluruh gradien dibatasi dengan max_norm = 1.0 agar tidak melebihi nilai tertentu, sehingga model tetap stabil selama proses pembelajaran.

- *Learning Rate Scheduling* dengan *ReduceLROnPlateau*

Model menggunakan scheduler ReduceLROnPlateau yang secara adaptif menurunkan *learning rate* apabila akurasi validasi tidak menunjukkan peningkatan. Hal ini penting untuk menjaga efisiensi konvergensi, terutama ketika model mulai memasuki *plateau* dalam pembelajaran.

- *Early Stopping* dan *Checkpointing Model*

Untuk menghindari pelatihan berlebihan (*overtraining*), diterapkan *early stopping* berdasarkan nilai akurasi validasi terbaik. Apabila akurasi validasi tidak meningkat dalam lima epoch berturut-turut, pelatihan akan dihentikan secara otomatis. Model terbaik berdasarkan validasi akan disimpan menggunakan "torch.save()" dan dimuat kembali setelah tahap *training* selesai untuk evaluasi akhir

F. *Evaluasi Kinerja Model*

Untuk menilai performa model klasifikasi Vision Transformer (ViT) secara objektif, penelitian ini menggunakan serangkaian metrik evaluasi standar yang umum diterapkan dalam tugas klasifikasi biner. Metrik utama yang digunakan meliputi akurasi, presisi, recall, dan F1-score, yang masing-masing memberikan perspektif berbeda terhadap kinerja model:

- Akurasi (*accuracy*) menggambarkan sejauh mana model mampu memprediksi dengan benar dibandingkan seluruh jumlah prediksi yang dilakukan.
- Presisi (*precision*) mengukur proporsi prediksi positif yang benar-benar tepat sasaran (*true positive*), sehingga relevan dalam konteks di mana kesalahan positif (*false positive*) perlu diminimalkan.
- *Recall* menunjukkan seberapa baik model mengenali semua instance positif aktual, yaitu sejauh mana true positive berhasil ditangkap dari seluruh populasi positif.
- *F1-score*, sebagai *harmonic mean* dari presisi dan recall, memberikan ukuran seimbang yang berguna ketika terjadi ketidakseimbangan antar kelas atau ketika penting untuk mengoptimalkan keduanya secara bersamaan.

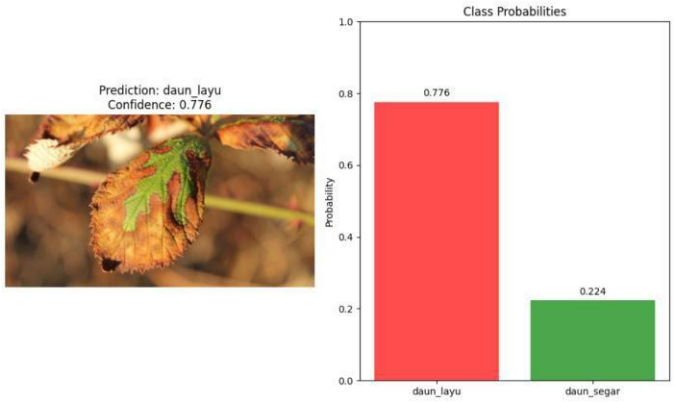
Implementasi perhitungan metrik-metrik ini dilakukan dengan memanfaatkan fungsi-fungsi dari *library* scikit-learn (sklearn.metrics), seperti accuracy_score, classification_report, dan confusion_matrix. Fungsi classification_report secara otomatis menghasilkan nilai presisi, recall, dan F1-score untuk masing-masing kelas secara terperinci, sehingga memudahkan interpretasi performa model terhadap masing-masing label daun segar dan daun layu.

Sementara itu, confusion matrix digunakan untuk memberikan gambaran visual tentang distribusi hasil prediksi terhadap label aktual. Matriks ini menyajikan jumlah instance yang diprediksi benar maupun salah untuk setiap kelas, sehingga dapat digunakan untuk menelusuri pola kesalahan yang terjadi, misalnya apakah model lebih sering salah mengklasifikasikan daun layu sebagai segar, atau sebaliknya. Seluruh hasil evaluasi kuantitatif akan disajikan dalam bentuk tabel ringkasan metrik serta visualisasi confusion matrix pada bagian Hasil dan Analisa

IV. HASIL DAN ANALISA

A. *Hasil Pelatihan Model*

Hasil evaluasi terhadap model Vision Transformer (ViT) yang diterapkan pada dataset citra daun menunjukkan bahwa proses pelatihan menghasilkan pola konvergensi yang stabil dan performa model yang relatif konsisten. Hal ini mengindikasikan bahwa model mampu belajar secara efektif dari data yang tersedia tanpa mengalami gejala overfitting yang signifikan. Stabilitas ini menjadi indikator penting bahwa pendekatan yang digunakan baik dari sisi arsitektur model maupun strategi pelatihan berjalan sesuai harapan.



Gambar 4. Hasil Testing

Model ViT yang digunakan merupakan model (*pre-trained*) pada dataset berskala besar seperti ImageNet-21k. Keuntungan utama dari pendekatan ini adalah kemampuan model dalam mengenali dan mengekstraksi representasi visual yang umum dijumpai dalam berbagai objek alami. Fitur-fitur visual ini, yang mencakup pola tekstur, warna, tepi, dan bentuk, ternyata dapat ditransfer dan diadaptasi dengan cukup baik untuk klasifikasi objek lain, yaitu membedakan antara daun segar dan daun layu.

Lebih jauh, kemampuan ViT dalam memanfaatkan *self-attention* untuk memahami relasi global antar bagian citra memungkinkan model menangkap perbedaan halus dalam

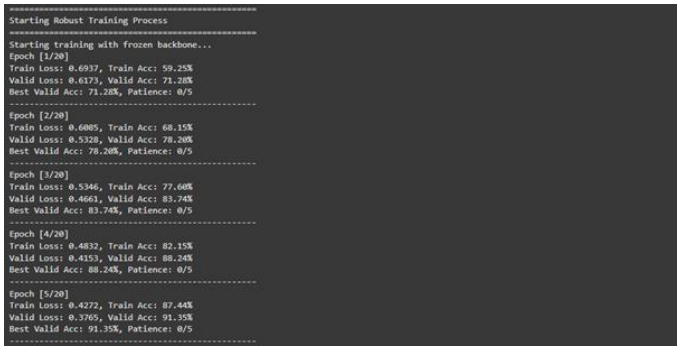
karakteristik visual daun, seperti perubahan warna, kerutan, atau tanda-tanda kekeringan yang mungkin sulit dideteksi secara konsisten oleh metode berbasis CNN tradisional atau bahkan oleh manusia tanpa pengalaman khusus.

B. Kinerja Klasifikasi

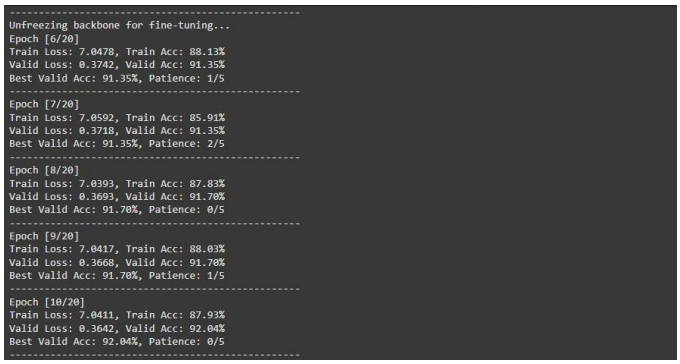
Hasil evaluasi model Vision Transformer pada dataset pengujian yang terdiri dari 3.000 citra daun menunjukkan kinerja yang memuaskan untuk tugas klasifikasi biner. Metrik-metrik evaluasi memberikan gambaran komprehensif tentang kemampuan model dalam mengidentifikasi kondisi daun secara akurat.

Tabel 1. Ringkasan Metrik Klasifikasi Model Transformer

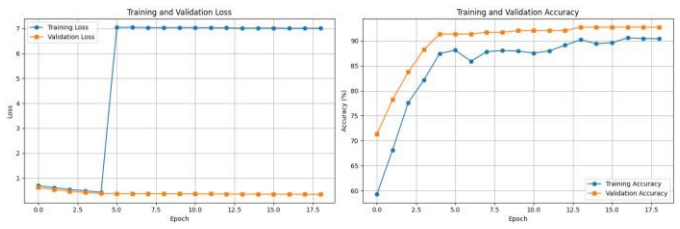
Metrik	Daun Segar	Daun Layu	Macro Average	Weighted Average
Presisi	0.84	0.97	0.90	0.90
Recall	0.97	0.80	0.89	0.89
F1-Score	0.90	0.88	0.89	0.89
Support	75	70	145	145



Gambar 5. Training Process Epoch 1 – 5



Gambar 6. Training Process Epoch 6 – 10 (Fine-Tuning)



Gambar 7. Training and Validation

- Tren Loss: Stabilitas dan Anomali Pelatihan

Pada grafik sebelah kiri, terlihat bahwa loss validasi menurun secara konsisten sejak awal pelatihan, mengindikasikan bahwa model semakin baik dalam mengklasifikasikan data yang belum pernah dilihat. Namun, terdapat anomali yang signifikan pada epoch ke-5, di mana training loss mengalami lonjakan tajam. Hal ini sejalan dengan strategi gradual unfreezing yang diterapkan dalam pelatihan, di mana backbone ViT dibuka untuk pelatihan penuh mulai epoch ke-5. Akibatnya, model memerlukan beberapa epoch untuk beradaptasi ulang dengan seluruh parameter yang kini dapat dilatih. Setelah itu, loss kembali stabil.

- Tren Akurasi: Konsistensi dan Konvergensi

Grafik kanan menunjukkan akurasi pelatihan dan validasi yang secara umum meningkat tajam pada lima epoch pertama dan kemudian mengalami konvergensi. Akurasi validasi bahkan sedikit lebih tinggi dibanding akurasi pelatihan sepanjang sisa epoch, sebuah indikasi bahwa model tidak mengalami overfitting, dan justru memiliki kemampuan generalisasi yang kuat.

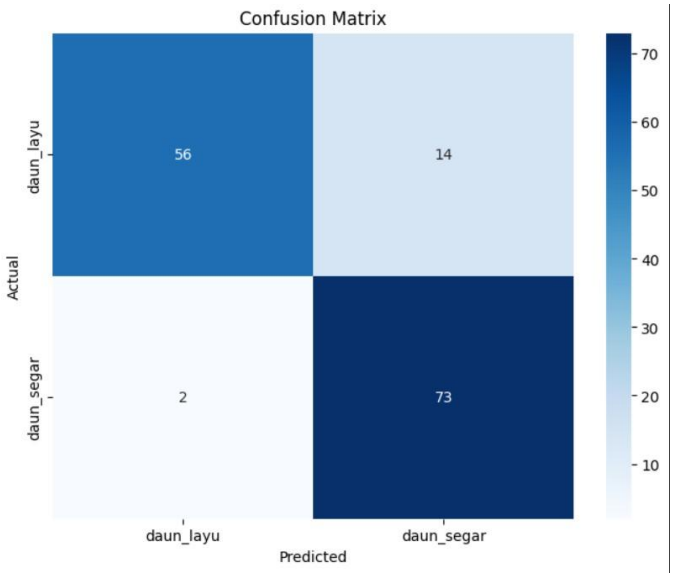
Puncak akurasi validasi mencapai lebih dari 91%, sementara akurasi pelatihan stabil di kisaran 88– 90%. Gap kecil ini dapat disebabkan oleh teknik regularisasi seperti label smoothing, dropout, dan early stopping, yang menjaga model agar tidak terlalu menyesuaikan diri pada data pelatihan.

- Interpretasi Kualitatif

Model ViT, yang dilatih dengan pendekatan transfer learning dari ImageNet, berhasil memanfaatkan representasi visual umum yang telah dipelajari sebelumnya untuk mengenali ciri khas daun segar dan layu. Ciri-ciri seperti perubahan warna, kekusaman, atau pola kerutan pada daun tampaknya berhasil dipelajari oleh model, sebagaimana tercermin dari akurasi validasi yang tinggi dan stabil.

C. Analisis Hasil

Hasil evaluasi yang ditampilkan dalam Gambar 7 (*Confusion Matrix*) dan Tabel 1 (Ringkasan Metrik Klasifikasi).



Gambar 7. Confusion Matrix

Hasil tersebut memberikan gambaran menyeluruh mengenai efektivitas model ViT dalam melakukan klasifikasi citra daun ke dalam dua kelas yaitu daun segar dan daun layu. Evaluasi ini tidak hanya menyoroti tingkat akurasi model secara keseluruhan, tetapi juga mengungkap dinamika performa model pada masing-masing kelas secara terpisah, termasuk potensi bias maupun kekuatan generalisasi.

Secara umum, model menunjukkan kinerja klasifikasi yang sangat baik, dengan skor F1 masing-masing sebesar 0.90 untuk daun segar dan 0.88 untuk daun layu. Tingginya skor F1 mengindikasikan bahwa model berhasil menjaga keseimbangan antara presisi dan recall dalam mengidentifikasi kedua jenis daun. Akurasi total dari pengujian mencerminkan bahwa mayoritas citra dapat diklasifikasikan dengan benar oleh model, yang mengindikasikan bahwa representasi fitur yang dipelajari oleh ViT cukup relevan terhadap domain visual daun. Berdasarkan visualisasi confusion matrix:

- Dari 70 citra daun layu, model berhasil mengklasifikasikan 56 dengan benar dan salah mengklasifikasikan 14 sebagai daun segar.
- Sementara itu, dari 75 citra daun segar, sebanyak 73 diklasifikasikan benar, dan hanya 2 citra yang salah klasifikasi sebagai layu.

Pola ini menunjukkan bahwa model lebih sensitif terhadap kelas "daun segar" (recall 0.97), namun sedikit kurang presisi dalam mendeteksi daun layu (recall 0.80). Hal ini bisa jadi disebabkan oleh karakteristik visual daun layu yang lebih bervariasi dan kurang konsisten, seperti terkait tingkat pelayuan, pencahayaan, atau kondisi latar belakang saat pengambilan gambar.

V. KESIMPULAN

Penelitian ini berhasil mendemonstrasikan efektivitas implementasi model Vision Transformer (ViT) dalam klasifikasi citra daun menjadi dua kategori utama yaitu daun segar dan daun layu. Dengan memanfaatkan model ViTForImageClassification dari *library* Hugging Face Transformers dan pendekatan transfer learning, sistem yang dikembangkan mampu mencapai akurasi keseluruhan sebesar 89% pada dataset pengujian yang terdiri dari 3.000 citra daun yang dikumpulkan secara langsung. Hasil ini menunjukkan bahwa ViT dapat mengenali dan membedakan kondisi visual daun secara akurat, bahkan dalam konteks variasi latar belakang, pencahayaan, dan kualitas citra yang umum dijumpai dalam kondisi lapangan.

Kontribusi utama dari penelitian ini mencakup tiga aspek:

- Pengembangan dataset lokal yang berimbang dan representatif terhadap kondisi nyata di lapangan
- Perancangan pipeline klasifikasi end-to-end yang mencakup tahap pra-pemrosesan, augmentasi, pelatihan, dan evaluasi model
- Analisis komprehensif kinerja model melalui metrik evaluasi yang mencakup akurasi, presisi, recall, F1- score, dan confusion matrix.

Hasil yang dicapai memberikan gambaran bahwa pendekatan berbasis deep learning, khususnya dengan arsitektur Vision Transformer, memiliki potensi besar untuk diterapkan dalam sistem monitoring tanaman otomatis. Model yang dihasilkan dapat menjadi pondasi awal bagi pengembangan aplikasi otomatisasi proses sortir daun atau hasil panen berdasarkan kondisi visual.

Keberhasilan ini sekaligus membuka ruang eksplorasi lebih lanjut dalam pengembangan sistem pertanian presisi berbasis citra, baik melalui peningkatan skala dataset pengayaan jenis objek tanaman yang diamati, maupun integrasi dengan perangkat IoT atau aplikasi mobile untuk implementasi lapangan secara real-time.

DAFTAR PUSTAKA

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput. Electron. Agric.*, vol. 147, pp. 70–90, 2018.

[3] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, p. 2674, 2018.

[4] J. Liang, D. Wang, and X. Ling, "Image classification for soybean and weeds based on ViT," in *J. Phys.: Conf. Ser.*, vol. 2002, no. 1, p. 012068, Aug. 2021.

[5] A. S. Nugroho and A. B. Witarto, "Penerapan Vision Transformer untuk Klasifikasi Citra Daun Tanaman Obat Indonesia," *J. Informatika*, vol. 16, no. 1, pp. 45–52, 2022.

[6] D. I. Patrício and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review," *Comput. Electron. Agric.*, vol. 153, pp. 69– 81, 2018.

[7] S. K. Sharma and D. K. Vishwakarma, "Classification of banana plant leaves based on nutrient deficiency using Vision Transformer," in *2024 5th Int. Conf. for Emerging Technol. (INCET)*, 2024, pp. 1–6.

[8] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019.

[9] R. P. Sari, A. Wibowo, and H. Suhartanto, "Implementasi Deep Learning untuk Klasifikasi Penyakit Daun Padi Menggunakan Convolutional Neural Network," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 4, pp. 735–742, 2021.

[10] A. Singh, B. Ganapathysubramanian, A. K. Singh, and S. Sarkar, "Machine learning for high-throughput stress phenotyping in plants," *Trends Plant Sci.*, vol. 21, no. 2, pp. 110–124, 2016.

[11] W. Ullah, K. Javed, M. A. Khan, F. Y. Alghayadh, M. W. Bhatt, I. S. Al Naimi, and I. Ofori, "Efficient identification and classification of apple leaf diseases using lightweight vision transformer (ViT)," *Discover Sustainability*, vol. 5, no. 1, p. 116, 2024.

[12] D. R. Wijaya, R. Sarno, and E. Zulaika, "Klasifikasi Citra Daun Tanaman Menggunakan Convolutional Neural Network (CNN) dan Support Vector Machine (SVM)," *J. Tek. ITS*, vol. 8, no. 2, pp. A169–A174, 2019