

**KLASIFIKASI PENYAKIT DAUN SIRIH MENGGUNAKAN
MODEL ARSITEKTUR *CONVOLUTIONAL NEURAL NETWORK*
DAN *VISION TRANSFORMER***

TUGAS AKHIR

Diajukan oleh:

Durri Munawar

220705037

Mahasiswa Fakultas Sains dan Teknologi

Program Studi Teknologi Informasi



**FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI AR-RANIRY
BANDA ACEH
2026 M/1448 H**

LEMBAR PERSETUJUAN

KLASIFIKASI PENYAKIT DAUN SIRIH MENGGUNAKAN MODEL ARSITEKTUR *CONVOLUTIONAL NEURAL NETWORK* DAN *VISION TRANSFORMER*

TUGAS AKHIR

Diajukan kepada Fakultas Sains dan Teknologi Universitas Islam Negeri Ar-Raniry
Banda Aceh Sebagai Salah Satu Beban Studi Memperoleh Gelar Sarjana (S1)
Pada Prodi Teknologi Informasi

Oleh:

DURRI MUNAWAR
NIM. 220705037

Mahasiswa Fakultas Sains dan Teknologi
Program Studi Teknologi Informasi

Disetujui Untuk di Munaqasyahkan Oleh:

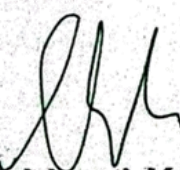
Pembimbing I,

Pembimbing II,


Dr. Hendri Ahmadian, S.Si., M.I.M.
NIP. 198301042014031002


Khairan AR, M.Kom.
NIP. 198607042014031001

Mengetahui,
Ketua Program Studi Teknologi Informasi


Malahayati, M.T.
NIP. 198301272015032003

LEMBAR PENGESAHAN

KLASIFIKASI PENYAKIT DAUN SIRIH MENGGUNAKAN MODEL ARSITEKTUR *CONVOLUTIONAL NEURAL NETWORK* DAN *VISION TRANSFORMER*

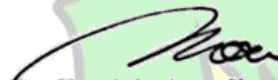
TUGAS AKHIR

Telah Diuji oleh Panitia Ujian Munaqasyah Tugas Akhir
Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh dan Dinyatakan Lulus
Serta Diterima Sebagai Salah Satu Beban Studi Program Sarjana (S1)
Dalam Program Studi Teknologi Informasi

Pada Hari/Tanggal: Senin, 10 Agustus 2026 M
26 Shafar 1448 H
Di Darussalam, Banda Aceh

Panitia Ujian Munaqasyah Tugas Akhir:

Ketua,


Dr. Hendri Ahmadian, M.I.M.
NIP. 198301042014031002

Sekretaris,


Khairan Ar, M.Kom.
NIP. 198607042014031001

Penguji I,


Malahayati, M.T.
NIP. 198301272015032003

Penguji II,


Fathiah, M.Eng.
NIP. 198606152019032010

Mengetahui:


Dekan Fakultas Sains dan Teknologi
UIN Ar-Raniry Banda Aceh

Prof. Dr. J. Muhammad Dirhamsyah, M.T., IPU
NIP. 196210021988111001

LEMBAR PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan di bawah ini:

Nama : Durri Munawar

NIM : 220705037

Program Studi : Teknologi Informasi

Fakultas : Sain dan Teknologi

Judul : Klasifikasi Penyakit Daun Sirih Menggunakan Model
Arsitektur *Convolutional Neural Network* Dan *Vision Transformer*

Dengan ini menyatakan bahwa dalam penulisan tugas akhir ini, saya:

1. Tidak menggunakan ide orang lain tanpa mampu mengembangkan dan mempertanggungjawabkan;
2. Tidak melakukan plagiasi terhadap naskah karya orang lain;
3. Tidak menggunakan karya orang lain tanpa menyebutkan sumber asli atau tanpa izin pemilik karya;
4. Tidak memanipulasi dan memalsukan data;
5. Mengerjakan sendiri karya ini dan mampu bertanggungjawab atas karya ini.

Bila dikemudian hari ada tuntutan dari pihak lain atas karya saya, dan telah melalui pembuktian yang dapat dipertanggungjawabkan dan ternyata memang ditemukan bukti bahwa saya telah melanggar pernyataan ini, maka saya siap dikenai sanksi berdasarkan aturan yang berlaku di Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh. Demikian pernyataan ini saya buat dengan sesungguhnya dan tanpa paksaan dari pihak manapun.

Banda Aceh, 11 Agustus 2026

Yang Menyatakan,



Durri Munawar

ABSTRAK

Nama : Durri Munawar
NIM : 22070503
Program Studi : Teknologi Informasi
Fakultas : Sain dan Teknologi
Judul : Klasifikasi Penyakit Daun Sirih Menggunakan Model
Arsitektur *Convolutional Neural Network* Dan *Vision Transformer*
Tanggal Sidang : 10 Agustus 2026 / 27 Safar 1448 H
Jumlah Halaman : 64
Pembimbing 1 : Dr. Hendri Ahmadian, M.I.M.
Pembimbing 2 : Khairan AR, M.Kom.

Tanaman sirih (*Piper betle L.*) merupakan komoditas bernilai tinggi yang rentan terhadap serangan patogen yang dapat menurunkan kualitas panen. Deteksi penyakit secara manual sering kali subjektif, memakan waktu, dan tidak efisien untuk lahan skala besar. Oleh karena itu, penelitian ini bertujuan untuk mengimplementasikan dan membandingkan performa dua arsitektur *deep learning* mutakhir, yaitu *Convolutional Neural Network* (ConvNeXt) dan *Vision Transformer* (DINOv2), dalam mengklasifikasikan kondisi penyakit pada daun sirih. Penelitian ini menggunakan 2.307 citra daun sirih yang dikategorikan ke dalam tiga kelas: *Healthy Leaf*, *Bacterial Leaf Disease*, dan *Dried Leaf*. Pelatihan model dilakukan melalui teknik *transfer learning* dengan melakukan *fine-tuning* pada parameter dasar arsitektur. Kinerja model dievaluasi secara kuantitatif menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Hasil pengujian menunjukkan bahwa arsitektur ConvNeXt secara signifikan mengungguli DINOv2 dengan capaian akurasi sebesar 99,03% berbanding 93,72%. Analisis metrik membuktikan ConvNeXt sangat stabil dengan keseimbangan *F1-Score* yang tinggi di seluruh kelas, sementara DINOv2 menunjukkan kelemahan *recall* pada pendeteksian infeksi bakteri berskala kecil. Penelitian ini menyimpulkan bahwa arsitektur CNN modern (ConvNeXt) lebih superior, akurat, dan direkomendasikan

untuk pengembangan sistem deteksi otomatis penyakit daun sirih dibandingkan *Vision Transformer*.

Kata kunci: Klasifikasi Citra, Daun Sirih, *Deep Learning*, ConvNeXt, DINOv2.



ABSTRACT

Name : Durri Munawar
Student Id : 220705037
Study Program : Information Technology
Faculty : Science and Technology
Title : Classification of Betel Leaf Diseases Using Convolutional Neural Network and Vision Transformer Architecture Models
Date : August 10, 2026 / 27 Safar 1448 H
Number Of Page : 64
Supervisor I : Dr. Hendri Ahmadian, M.I.M.
Supervisor II : Khairan AR, M.Kom.

Betel leaf (*Piper betle L.*) is a valuable commodity susceptible to various pathogen attacks that can degrade crop quality. Manual disease detection is often subjective, time-consuming, and inefficient for large-scale farming. Therefore, this research aims to implement and compare the performance of two state-of-the-art deep learning architectures, namely Convolutional Neural Network (ConvNeXt) and Vision Transformer (DINOv2), in classifying disease conditions on betel leaves. This study utilized 2,307 betel leaf images categorized into three classes: Healthy Leaf, Bacterial Leaf Disease, and Dried Leaf. Model training was conducted using transfer learning by fine-tuning the foundational architecture parameters. Model performance was evaluated quantitatively using Accuracy, Precision, Recall, and F1-Score metrics. The test results indicate that the ConvNeXt architecture significantly outperforms DINOv2, achieving an accuracy of 99.03% compared to 93.72%. Metric analysis proves that ConvNeXt is highly stable with a high F1-Score balance across all classes, whereas DINOv2 exhibits recall weaknesses in detecting small-scale bacterial infections. This research concludes that the modern CNN architecture (ConvNeXt) is superior, more accurate, and recommended for the development of automated betel leaf disease detection systems compared to the Vision Transformer.

Keywords: Image Classification, Betel Leaf, Deep Learning, ConvNeXt, DINOv2.

KATA PENGANTAR

Alhamdulillahahirabbil'aalamiin. Segala Puji bagi Allah SWT Tuhan yang Maha Esa senantiasa panjatkan kehadirat Allah SWT, Allah SWT yang telah melimpahkan rahmat, hidayah-Nya, sehingga penulis dapat menyelesaikan penyusunan tugas akhir dengan judul “Klasifikasi Penyakit Daun Sirih menggunakan Model Arsitektur *Convolutional Neural Network* dan *Vision Transformer*”.

Shalawat beriringan dengan salam senantiasa tercurahkan kepada junjungan alam, Nabi besar Muhammad SAW, serta keluarga dan sahabat beliau yang telah membantu beliau dalam menjalankan perintah Allah SWT, yaitu memperluas ilmu pengetahuan dan cahaya iman untuk umat manusia.

Penelitian ini disusun sebagai salah satu tahapan prasyarat dalam pengerjaan tugas akhir di bawah asuhan Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Ar-raniry Banda Aceh.

Penulis menyadari sepenuhnya bahwa penyelesaian tugas akhir ini tidak terlepas dari bantuan, bimbingan, arahan, dan dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Kepada Orang tua dan Keluarga tercinta yang selalu memberikan Doa, kasih sayang, kepedulian dan dukungan yang besar dari segi apapun.
2. Bapak Prof. Dr. Ir. M. Dirhamsyah, M.T., IPU selaku Dekan Fakultas Sains dan Teknologi UIN Ar-Raniry.
3. Ibu Malahayati, M.T., yang menjabat sebagai Ketua Program Studi Teknologi Informasi sekaligus pembimbing yang telah memberikan masukan, bimbingan dan arahan selama proses penyusunan proposal ini.
4. Bapak Nazaruddin Ahmad, M.T. selaku Dosen Pembimbing Akademik. (PA) yang telah memberikan bimbingan dan arahan selama masa.

5. Bu Cut Ida Rahmadiana, S. Si., selaku Staf dan pegawai di Program Studi Teknologi Informasi yang telah banyak membantu, bimbingan, dan kerja sama yang baik.
6. Seluruh Dosen dan Staf Akademik Program Studi Teknologi Informasi yang telah memberikan ilmu, wawasan, dan pelayanan administrasi selama ini.
7. Penghargaan sebesar besarnya kepada Sahabat, Silva Yanti dan kawan seperjuangan di Program Studi Teknologi Informasi angkatan 2022 atas segala bentuk energi positif, dukungan, dan motivasi yang senantiasa mengiringi langkah penulis dalam merampungkan proposal ini.

Penulis menyadari bahwa tugas akhir ini masih jauh dari kata sempurna dan masih memiliki banyak kekurangan, baik dari segi penyusunan bahasa maupun keilmuan, mengingat keterbatasan pengetahuan yang penulis miliki. Oleh karena itu, kritik dan saran yang membangun sangat penulis harapkan demi perbaikan dan pengembangan di masa yang akan datang.

Akhir kata, penulis berharap semoga Allah SWT senantiasa membalas segala kebaikan semua pihak yang telah membantu, dan semoga tugas akhir ini dapat memberikan manfaat yang nyata, khususnya bagi perkembangan bidang Teknologi Informasi dan masyarakat luas pada umumnya. Aamiin Yaa Rabbal 'Aalamiin.

Banda Aceh, Agustus 2026

Durri Munawar

DAFTAR ISI

LEMBAR PERSETUJUAN	i
LEMBAR PENGESAHAN.....	ii
LEMBAR PERNYATAAN KEASLIAN TUGAS AKHIR	iii
ABSTRAK.....	iv
ABSTRACT	vi
KATA PENGANTAR	vii
DAFTAR ISI.....	ix
DAFTAR GAMBAR.....	xii
DAFTAR TABEL.....	xiii
BAB I PENDAHULUAN	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Tujuan Penelitian	3
1.4 Manfaat Penelitian	3
1.5 Batasan Masalah	3
BAB II TINJAUAN PUSTAKA.....	5
2.1 Penelitian Terdahulu	5
2.2 Landasan Teori	8
2.2.1 Penyakit Daun Sirih	8
2.2.2 Computer Vision.....	9
2.2.3 <i>Machine Learning</i>	9
2.2.4 Supervised Learning	10

2.2.5 Deep learning	10
2.2.6 Convolutional Neural Networks	11
2.2.7 ConvNeXt (Modernized Convolutional Neural Network).....	13
2.2.8 Transfer Learning.....	16
2.2.9 Transformer dan Self-Attention	16
2.2.10 Vision Transformer	17
2.2.11 DINOv2 (Distillation with NO labels).....	20
2.2.12 Evaluasi Performa	21
BAB III METODOLOGI PENELITIAN.....	23
3.1 Jenis Penelitian	23
3.2 Tahapan Penelitian.....	23
3.2.1 Pengumpulan Data	24
3.2.2 Persiapan dan Pengolahan Data	25
3.2.3 Perancangan Arsitektur Model.....	26
3.2.4 Pelatihan Model	29
3.3 Bahan dan Alat Penelitian.....	30
BAB IV HASIL DAN PEMBAHASAN.....	32
4.1 Skenario Pengujian	32
4.1.1 Persiapan dan Pembagian Dataset.....	32
4.1.2 Hyperparameter Model	34
4.2 Hasil pengujian	35
4.2.1 Hasil Pelatihan ConvNeXt	35
4.2.2 Hasil Pelatihan DINOv2	36
4.2.3 Evaluasi Metrik Kinerja	38

4.3 Pembahasan	42
BAB V KESIMPULAN DAN SARAN	44
5.1 Kesimpulan	44
5.2 Saran	45
DAFTAR PUSTAKA	47



DAFTAR GAMBAR

Gambar 2.1 Arsitektur Convolutional Neural Network	12
Gambar 2.2 Grafik Benchmark CNN dan ViT	14
Gambar 2. 3 Arsitektur ConvNeXt.....	15
Gambar 2.4 Arsitektur Vision Transforme.....	17
Gambar 3.1 Alur Penelitian	23
Gambar 3.2 Arsitektur Perancangan ConvNeXt pada Daun Sirih	27
Gambar 3.3 Arsitektur Perancangan DINOv2 pada Daun Sirih.....	28
Gambar 4. 1 Log metrik pelatihan per-epoch pada ConvNeXt.....	35
Gambar 4. 2 Grafik Konvergensi Loss dan Akurasi Validasi pada ConvNeXt	36
Gambar 4.3 Log metrik pelatihan per-epoch pada DINOv2	37
Gambar 4.4 Grafik Konvergensi Loss dan Akurasi Validasi pada DINOv2.....	38
Gambar 4.5 Confusion Matrix ConvNeXt dan DINOv2	41



DAFTAR TABEL

Tabel 2.1 Penelitian Terdahulu.....	5
Tabel 2.2 Varian Model Vision Transformer	19
Tabel 4.1 Rincian Pembagian Dataset Daun Sirih	33
Tabel 4.2 Konfigurasi Hyperparameter Pelatihan	34
Tabel 4.3 Classification Report Model ConvNeXt	38
Tabel 4.4 Classification Report Model DINOv2.....	39



BAB I

PENDAHULUAN

1.1 Latar Belakang

Tanaman sirih (*Piper betle L.*) merupakan salah satu komoditas berharga yang memiliki nilai ekonomi dan farmakologis tinggi di masyarakat. Sebagai tanaman penghasil minyak atsiri dan bahan baku obat tradisional, kualitas fisik daun sirih menjadi indikator utama dalam menentukan nilai jualnya di pasaran. Sayangnya, budidaya tanaman ini sangat rentan terhadap serangan berbagai patogen yang memicu penyakit seperti bercak daun (*leaf spot*) maupun infeksi bakteri. Serangan penyakit ini, jika tidak dideteksi dan ditangani dengan cepat, dapat merusak struktur daun dan menurunkan kualitas panen secara drastis, yang berdampak langsung pada kerugian ekonomi bagi para petani (Ahmed et al., 2021).

Hingga saat ini, proses identifikasi penyakit pada daun sirih di tingkat petani umumnya masih mengandalkan pengamatan visual secara manual. Metode konvensional ini memiliki banyak keterbatasan karena sangat bergantung pada pengalaman serta subjektivitas pengamat, sehingga rentan terhadap kesalahan diagnosis (*human error*). Selain itu, deteksi secara manual memakan waktu yang lama dan sangat tidak efisien apabila diterapkan pada lahan perkebunan berskala besar. Oleh karena itu, diperlukan sebuah sistem otomatisasi berbasis komputasi yang mampu mendiagnosis kondisi tanaman secara cepat, objektif, dan presisi (Jackulin & Murugavalli, 2022).

Perkembangan teknologi kecerdasan buatan, khususnya pada bidang *Deep learning*, telah membuka peluang besar dalam menyelesaikan permasalahan klasifikasi citra di sektor pertanian. Algoritma *Convolutional Neural Network* (CNN) selama bertahun-tahun menjadi standar utama yang diandalkan karena kemampuannya yang luar biasa dalam mengekstraksi fitur-fitur lokal, seperti warna, pola garis, dan tekstur bercak penyakit pada daun. Namun, arsitektur CNN memiliki keterbatasan pada *receptive field* atau area tangkapan visualnya, sehingga jaringan ini sering kali kurang

optimal dalam memahami konteks spasial gambar secara keseluruhan (global) (Urrea & Vélez, 2025).

Sebagai alternatif untuk mengatasi keterbatasan tersebut, arsitektur *Vision Transformer* (ViT) kini mulai diadaptasi secara luas dalam tugas visi komputer. Berbeda dengan CNN yang memproses gambar melalui operasi konvolusi, ViT membagi gambar menjadi kumpulan *patch* kecil dan menggunakan mekanisme *self-attention* untuk menangkap hubungan jarak jauh antar bagian gambar. Kemampuan ini membuat ViT sangat tangguh dalam memahami konteks global citra. Meski demikian, ViT murni seringkali membutuhkan ukuran dataset yang sangat masif untuk bisa beradaptasi secara optimal, dan terkadang kurang peka dalam menangkap detail fitur lokal berskala kecil pada data yang terbatas (Y. Wang et al., 2025).

Mengingat kedua arsitektur tersebut memiliki paradigma pemrosesan fitur yang berbeda di mana CNN unggul dalam fitur lokal dan ViT unggul dalam fitur global maka penting untuk mengevaluasi efektivitas masing-masing model dalam mendeteksi anomali pada daun sirih. Hingga saat ini, pemilihan model yang paling optimal antara arsitektur berbasis konvolusi dan arsitektur berbasis attention untuk kasus klasifikasi penyakit tanaman sirih masih menjadi tantangan penelitian yang krusial. Pemahaman mendalam mengenai bagaimana masing-masing arsitektur ini menangani variasi tekstur dan pola penyebaran penyakit sangat diperlukan untuk menentukan standar akurasi terbaik (Jackulin & Murugavalli, 2022).

Penelitian ini berfokus pada studi komparatif antara model *Deep learning* berbasis CNN dan ViT untuk klasifikasi tiga kategori kondisi daun sirih, yaitu Daun Sehat, Daun Tkelaerkena Bakteri dan Daun Kering . Pendekatan ini bertujuan untuk mengevaluasi secara mendalam keandalan kinerja masing-masing arsitektur dalam mengekstraksi karakteristik unik dari setiap penyakit, guna memberikan rekomendasi model yang paling presisi dan efisien untuk diimplementasikan di sektor pertanian.

1.2 Rumusan Masalah

Dari penyesuaian latar belakang, maka dapat dirumuskan permasalahan untuk penelitian ini sebagai berikut :

1. Bagaimana penerapan model *deep learning Convolutional Neural Network* (CNN) dan Vision Transformer (ViT) dalam melakukan klasifikasi penyakit daun sirih berdasarkan citra daun?
2. Sejauh mana performa model *Convolutional Neural Network* (CNN) dan Vision Transformer (ViT) mampu memberikan hasil klasifikasi yang akurat berdasarkan metrik evaluasi ?

1.3 Tujuan Penelitian

Berdasarkan Rumusan masalah sebelumnya, tujuan penelitian yang dapat diangkat adalah sebagai berikut :

1. Mengimplementasikan model *Convolutional Neural Network* (CNN) dan Vision Transformer (ViT) untuk klasifikasi penyakit daun sirih.
2. Menganalisis dan membandingkan performa model CNN dan Vision Transformer (ViT) menggunakan metrik evaluasi.

1.4 Manfaat Penelitian

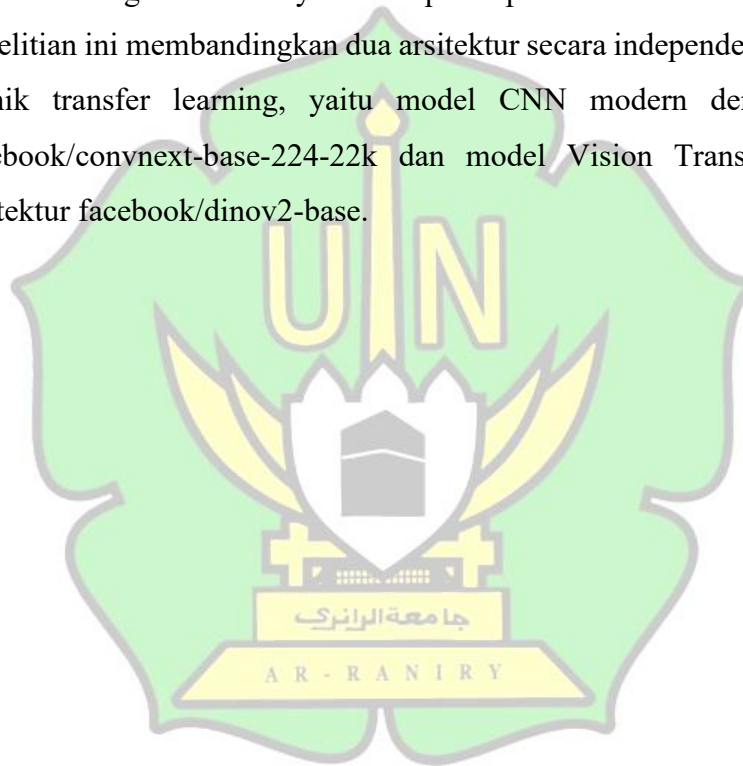
Manfaat yang diharapkan pada penelitian ini adalah sebagai berikut :

1. Penelitian ini diharapkan dapat memperkaya kajian mengenai teknik ekstraksi fitur lokal dan global guna meningkatkan akurasi klasifikasi citra penyakit tanaman, khususnya sebagai referensi dalam membandingkan keandalan arsitektur CNN dan ViT untuk tugas visi komputer di sektor pertanian.
2. Menjadi landasan pengembangan alat bantu deteksi penyakit daun sirih yang cepat dan akurat bagi para petani maupun masyarakat untuk ke depannya.

1.5 Batasan Masalah

Guna menjaga fokus penelitian dan memastikan tercapainya tujuan yang telah ditetapkan, maka penelitian ini dibatasi pada hal-hal sebagai berikut:

1. Penelitian ini hanya berfokus pada klasifikasi daun tunggal tanaman sirih (*Piper betle .L*) ke dalam tiga kategori klasifikasi yaitu: Daun Sehat, Daun terkena Bakteri dan Daun Kering.
2. Model yang digunakan terdiri dari dua arsitektur deep learning yang diuji secara terpisah, yaitu ConvNeXt (CNN) dan DINOv2 (Vision Transformer).
3. Dataset yang digunakan terdiri dari data sekunder (repositori publik seperti Kaggle) dan data primer yang diambil menggunakan perangkat seluler
4. Output penelitian dibatasi pada pembangunan model yang telah teruji serta tervalidasi tingkat akurasinya serta laporan performa model tersebut.
5. Penelitian ini membandingkan dua arsitektur secara independen menggunakan teknik transfer learning, yaitu model CNN modern dengan arsitektur facebook/convnext-base-224-22k dan model Vision Transformer dengan arsitektur facebook/dinov2-base.



BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian terdahulu disajikan untuk memberikan gambaran dan dasar pemahaman mengenai metode dan pendekatan yang telah digunakan pada penelitian sebelumnya. Kajian ini membantu penulis mengetahui kelebihan dan keterbatasan penelitian terdahulu sehingga dapat menjadi acuan dalam menentukan metode serta kontribusi kebaruan pada penelitian yang dilakukan. Beberapa ringkasan penelitian terdahulu disajikan pada table 2.1.

Tabel 2.1 Penelitian Terdahulu

No	Peneliti	Judul Penelitian	Fokus & Metode	Hasil
1	(Hridoy et al., 2022)	Deep Neural Networks-Based Recognition of Betel Plant Diseases by Leaf Image Classification	Fokus: Klasifikasi penyakit daun sirih menggunakan <i>Deep Learning</i> ; Menggunakan arsitektur Convolutional Neural Network, yaitu ResNet-50 pada 10.662 dataset citra.	Arsitektur ResNet-50 mencapai akurasi sebesar 86,2%.
2	(Bhowmik et al., 2024)	A customised vision transformer for accurate detection and classification of Java Plum leaf disease	Fokus: Deteksi penyakit daun jambang (<i>Java Plum</i>). Metode: Menggunakan Vision Transformer (ViT) yang dikustomisasi.	Model ViT kustom ini berhasil meraih akurasi sebesar 97,51%. Membuktikan ViT sangat

				tanggguh untuk mendeteksi penyakit tanaman.
3	(Hussien et al., 2025)	Explainable Self-Supervised Learning for Medical Image Diagnosis Based on DINO V2 Model and Semantic Search	Fokus: Klasifikasi citra medis menggunakan Self-Supervised Learning. Metode: Membandingkan DINOv2, BYOL, SimCLR, MoCo-v3, Barlow Twins, VGG19, ResNet152, dan EfficientNet-B7 pada empat dataset medis (Lung Cancer, Brain Tumor, Leukemia, dan Retina Disease). Dataset Retina terdiri dari sekitar 4.000 citra dengan 4 kelas.	DINOv2 memperoleh akurasi tertinggi pada dataset Retina Disease, yaitu 95%, mengungguli seluruh model pembanding baik supervised maupun self-supervised.
4	(Brondolo & Beaussant, 2025)	DINOv2 Rocks Geological Image Analysis: Classification, Segmentation, and Interpretability	Fokus: Klasifikasi citra geologi menggunakan Self-Supervised Vision Transformer. Metode: Menggunakan DINOv2 (Small, Base, Large) pada	Model DINOv2 mencapai akurasi klasifikasi hingga 100% menggunakan

			3.000 citra CT batuan yang terdiri dari 10 kelas.	classifier k-NN pada sebagian besar skenario pengujian.
5	(X. Wang et al., 2023)	ECA-ConvNeXt: A Rice Leaf Disease Identification Model Based on ConvNeXt	Fokus: Identifikasi enam kategori penyakit khas daun padi dan daun sehat. Metode: Mengusulkan model ECA-ConvNeXt yang berbasis pada jaringan ConvNeXt dengan menambahkan modul Efficient Channel Attention (ECA) serta menerapkan transfer learning.	Model usulan mencapai akurasi 94,82%, presisi 94,47%, recall 94,31%, dan F1-Score 94,33%. Model ConvNeXt ini terbukti mampu melampaui kinerja model klasifikasi State-of-the-Art lainnya seperti ResNet-34 dan ResNet-50.
6	(Murugesan et al., 2025)	Robust Multiclass Classification of Crop Leaf	Fokus: Klasifikasi multikelas penyakit daun tanaman menggunakan	Model Hybrid ConvNet-ViT memperoleh

		Diseases Using Hybrid Deep Learning and Grad-CAM Interpretability	Deep Learning. Metode: Membandingkan EfficientNetV2, ConvNeXt, Swin Transformer, Vision Transformer (ViT), dan Hybrid ConvNet-ViT pada 12.865 citra dengan 9 kelas penyakit daun.	akurasi tertinggi 99,29%, sedangkan ConvNeXt dan ViT masing-masing memperoleh akurasi 98,92%.
--	--	---	---	---

Berdasarkan tinjauan penelitian terdahulu pada Tabel 2.1, sebagian besar studi mengenai klasifikasi penyakit daun sirih masih terbatas pada pemanfaatan arsitektur *Convolutional Neural Network* konvensional. Di sisi lain, implementasi arsitektur mutakhir berbasis *self-supervised learning* seperti DINOv2 lebih banyak dieksplorasi pada domain medis dan citra satelit. Oleh karena itu, keterbaruan dari penelitian ini terletak pada studi komparatif secara langsung dan independen antara arsitektur CNN termmodern, yaitu ConvNeXt, dengan *Vision Transformer* berbasis *self-supervised learning*, yaitu DINOv2, secara spesifik untuk kasus pengenalan anomali daun sirih.

2.2 Landasan Teori

2.2.1 Penyakit Daun Sirih

Penelitian yang dilakukan oleh (Utami et al., 2022) menyoroti bahwa infeksi penyakit seperti busuk daun (leaf rot), hawar bakteri (bacterial leaf spot), dan bercak daun (leaf spot) merupakan ancaman utama yang memicu kerugian ekonomi signifikan bagi petani. Lebih lanjut, studi-studi tersebut menegaskan bahwa keterlambatan dalam mendeteksi anomali visual pada anatomi daun seperti munculnya lesi nekrotik berwarna gelap atau degradasi klorofil yang cepat sering kali berujung pada gagal panen akibat penyebaran patogen yang sangat masif dan sulit dikendalikan di area perkebunan.

Setiap jenis infeksi pada daun sirih secara alamiah akan memunculkan degradasi warna, perubahan tekstur, dan pola batas lesi yang spesifik. Kompleksitas dan kemiripan gejala visual antar penyakit ini menuntut sebuah pendekatan komputasi yang tangguh. Oleh karena itu, variasi visual dari penyakit daun sirih ini mendasari urgensi penggunaan algoritma pengenalan pola tingkat tinggi yang tidak hanya teliti dalam mengekstraksi detail bercak terkecil, tetapi juga mampu menginterpretasikan kondisi spasial daun secara utuh dan menyeluruh.

2.2.2 Computer Vision

Computer Vision merupakan bidang yang memungkinkan mesin untuk "Melihat" dengan menggunakan kamera dan komputer sebagai pengganti mata manusia untuk mengidentifikasi, melacak, dan mengukur objek sebelum diproses lebih lanjut secara digital. Kemajuan teknologi di bidang *Computer Vision* telah memberikan berbagai solusi inovatif dalam berbagai sektor, termasuk pertanian. (Pakiding et al., 2025) Proses evaluasi data visual ini melibatkan teknik ekstraksi fitur yang mengubah piksel gambar menjadi data digital yang bermakna, sehingga komputer mampu memahami karakteristik unik dari sebuah objek secara otomatis.

Dalam pertanian modern, teknologi ini jadi "mata otomatis" yang bantu petani pantau kesehatan tanaman tanpa harus keliling satu-persatu. Cukup lewat foto daun, sistem bisa mendeteksi perubahan kecil seperti bercak, tekstur aneh, atau warna yang berubah sebagai tanda awal adanya penyakit. Karena sistem ini pintar dan bekerja secara langsung (*real-time*), petani bisa langsung tahu jenis penyakitnya dan cepat mengambil tindakan supaya tanaman tidak rusak atau gagal panen.

2.2.3 Machine Learning

Machine Learning adalah cabang dari *artificial intelligence* yang memungkinkan sistem komputer untuk belajar dari data, mengenali pola, dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit untuk setiap aturan fiksinya (Erda et al., 2025). Konsep dasar *Machine Learning* didasarkan pada perubahan pendekatan dari pemrograman konvensional yang bersifat kaku menjadi sistem yang mampu belajar secara mandiri dari data. Pada pemrograman tradisional, pengembang harus

menentukan aturan secara eksplisit, misalnya dengan menetapkan kondisi tertentu untuk mengambil keputusan. Sebaliknya, dalam *Machine Learning*, sistem diberikan sejumlah besar data, seperti kumpulan citra daun, untuk dipelajari. Melalui proses pembelajaran ini, model secara otomatis mengenali pola dan karakteristik penting dalam data, termasuk detail yang sulit diamati secara langsung oleh manusia, seperti variasi tekstur atau perubahan warna yang halus. Dengan kemampuan tersebut, sistem dapat menghasilkan keputusan yang lebih akurat dan mampu beradaptasi ketika dihadapkan pada data baru yang belum pernah dipelajari sebelumnya.

Machine Learning memiliki metode yang membantu dalam memudahkan belajar secara otomatis berdasarkan pengalaman dari data secara otomatis antara lain *Supervised learning* yang mempelajari hubungan dari data yang sudah diberi label, *Unsupervised learning* yang menemukan pola dalam data tanpa label, serta varian lain seperti *semi-supervised* dan *reinforcement learning* yang menyesuaikan dengan kebutuhan problem yang dipecahkan (Alnuaimi & Albaldawi, 2024).

2.2.4 Supervised Learning

Supervised learning merupakan metode pembelajaran *Machine Learning* dimana di mana sebuah algoritma cerdas dilatih menggunakan sekumpulan dataset yang telah memiliki label. Menurut (Hu et al., 2020) *Supervised Learning* umumnya diterapkan pada dua jenis permasalahan utama, yaitu klasifikasi untuk memprediksi kelas atau kategori tertentu dan regresi untuk memprediksi nilai kontinu. Dibandingkan dengan metode pembelajaran lainnya, seperti *unsupervised learning* yang tidak menggunakan label dan *reinforcement learning* yang mengandalkan mekanisme penghargaan (*reward*), supervised learning lebih sesuai digunakan ketika tersedia data yang telah diberi label sehingga proses pembelajaran dapat berlangsung secara terarah dan menghasilkan tingkat akurasi yang tinggi.

2.2.5 Deep learning

Deep learning adalah cabang dari *Machine Learning* yang menggunakan jaringan saraf tiruan berlapis untuk secara otomatis mempelajari pola dan representasi kompleks dari data dalam jumlah besar (Alzubaidi et al., 2021). *Deep learning* sendiri

memiliki kemampuan tersendiri yaitu mempelajari dan mengekstraksi fitur secara otomatis dari data mentah tanpa perlu perancangan fitur manual.

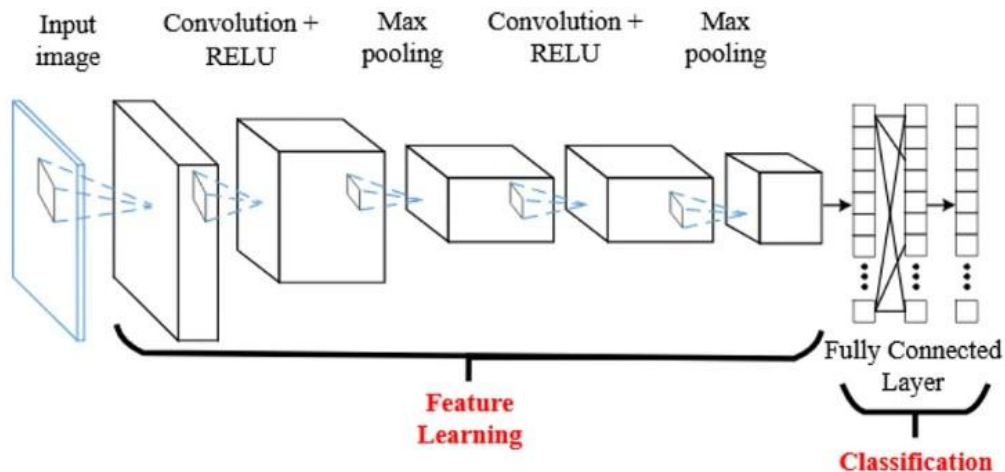
Pemrosesan *deep learning* pada data gambar meliputi pengolahan piksel gambar melalui jaringan saraf berlapis untuk mempelajari pola visual, mulai dari fitur sederhana seperti tepi hingga fitur kompleks seperti tekstur dan bercak penyakit. Fitur-fitur tersebut direpresentasikan secara numerik sehingga model dapat membedakan kelas citra berdasarkan pola visualnya. Selama pelatihan, bobot jaringan disesuaikan menggunakan data berlabel untuk meminimalkan kesalahan, dan hasil pembelajaran ini digunakan untuk mengklasifikasikan penyakit pada citra daun yang belum pernah dilihat sebelumnya (Hassan et al., 2022)

2.2.6 Convolutional Neural Networks

Convolutional Neural Network (CNN) adalah salah satu model *deep learning* yang banyak digunakan dalam visi komputer karena kemampuannya mempelajari dan mengekstraksi fitur dari data citra secara otomatis untuk berbagai tugas seperti klasifikasi, deteksi objek, dan segmentasi, serta terus berevolusi untuk menghadapi tantangan yang semakin kompleks dalam pengolahan citra (Zhao et al., 2024). Keunggulan utama CNN terletak pada penggunaan filter konvolusi dan *parameter sharing*.

Filter konvolusi memungkinkan model mengekstraksi fitur visual secara otomatis dan bertahap, mulai dari ciri sederhana seperti tepi dan warna hingga pola kompleks seperti tekstur dan bercak penyakit pada daun sirih. Sementara itu, *parameter sharing* membuat satu filter yang sama digunakan di seluruh bagian citra, sehingga pola penyakit dapat dikenali meskipun muncul di lokasi yang berbeda. Kombinasi keduanya menjadikan CNN efisien, stabil, dan efektif dalam melakukan klasifikasi citra berbasis pola visual.

Secara struktural, arsitektur dasar dari sebuah jaringan CNN terbagi menjadi dua tahapan utama, yaitu fase ekstraksi fitur (*Feature Learning*) dan fase klasifikasi (*Classification*), sebagaimana diilustrasikan pada Gambar 2.1 berikut:



Gambar 2.1 Arsitektur Convolutional Neural Network
Sumber: (Shehan et al., 2023)

Berdasarkan Gambar 2.1, alur kerja arsitektur CNN dalam memproses sebuah citra dapat diuraikan ke dalam dua tahap sebagai berikut:

1. Tahap Pembelajaran Fitur (*Feature Learning*).

Tahap ini merupakan tulang punggung arsitektur CNN yang bertugas menerjemahkan citra piksel mentah menjadi representasi fitur matematis yang kaya informasi. Tahapan ini beroperasi secara berulang melalui beberapa lapisan:

- a. *Input Image*: Matriks piksel dari citra masukan (misalnya citra daun sirih) diumpankan ke dalam jaringan.
- b. *Convolution + ReLU*: Lapisan konvolusi (*Convolutional Layer*) menggunakan filter (kernel) yang bergeser ke seluruh bagian gambar untuk memindai dan mendeteksi fitur lokal, seperti garis batas daun atau tekstur bercak penyakit. Hasil konvolusi ini kemudian dilewatkan pada fungsi aktivasi ReLU (*Rectified Linear Unit*). Fungsi aktivasi ini membuang semua nilai piksel negatif dengan mengubahnya menjadi 0, yang bertujuan untuk mempercepat proses komputasi dan menonjolkan fitur-fitur yang relevan.

- c. *Max Pooling*: Lapisan ini berfungsi untuk mereduksi dimensi ukuran spasial citra (*downsampling*) dengan cara hanya mengambil nilai piksel maksimal pada suatu area (filter) tertentu. Tujuannya adalah untuk meringankan beban komputasi komputer, meminimalkan risiko *overfitting*, dan membuat model kebal terhadap perubahan posisi objek di dalam gambar.
- d. Proses dari blok *Convolution + ReLU* dan *Max Pooling* ini umumnya diulang beberapa kali. Semakin dalam lapisannya, model akan mengenali fitur hierarkis yang semakin kompleks.

2. Tahap Klasifikasi (*Classification*).

Setelah gambar selesai dipindai dan direduksi ukurannya menjadi matriks kecil yang sangat padat fitur (hasil *pooling* terakhir), proses selanjutnya diambil alih oleh tahap klasifikasi:

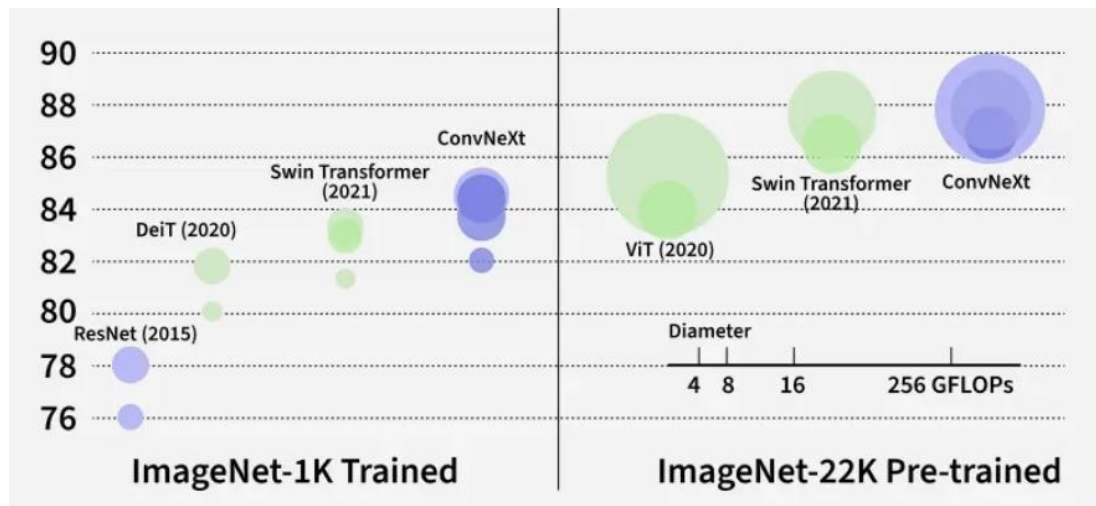
- a. *Flattening*: Meskipun tidak dituliskan secara eksplisit pada bagan, panah yang menghubungkan blok 3D terakhir menuju kumpulan titik saraf 1D mewakili proses perataan (*flattening*). Proses ini mengubah matriks fitur multidimensi menjadi satu vektor atau barisan angka linear.
- b. *Fully Connected Layer*: Vektor satu dimensi tersebut kemudian dimasukkan ke dalam jaringan saraf tiruan konvensional (*Fully Connected Layer*). Saraf-saraf pada lapisan ini saling terhubung secara penuh untuk mempelajari kombinasi akhir dari seluruh fitur visual. Pada ujung lapisan ini, model akan menghitung probabilitas matematis untuk menentukan keputusan akhirnya (misalnya memprediksi apakah citra tersebut masuk ke dalam kelas *Healthy*, *Bacterial*, atau *Dried*).

2.2.7 ConvNeXt (Modernized Convolutional Neural Network)

Meskipun arsitektur *Vision Transformer* (ViT) berkembang pesat dalam tugas visi komputer, arsitektur *Convolutional Neural Network* (CNN) mengalami modernisasi yang sangat signifikan. ConvNeXt merupakan arsitektur CNN modern

yang diperkenalkan oleh (Liu et al., 2022) dengan merancang ulang struktur jaringan ResNet standar agar mengadopsi prinsip-prinsip desain *Vision Transformer* tanpa meninggalkan operasi dasar konvolusi.

Sebagai landasan pemilihan model pada penelitian ini, keunggulan arsitektur ConvNeXt dibandingkan model *State-of-the-Art* lainnya dapat dilihat secara empiris melalui grafik *benchmark* resmi pada Gambar 2.2 berikut:

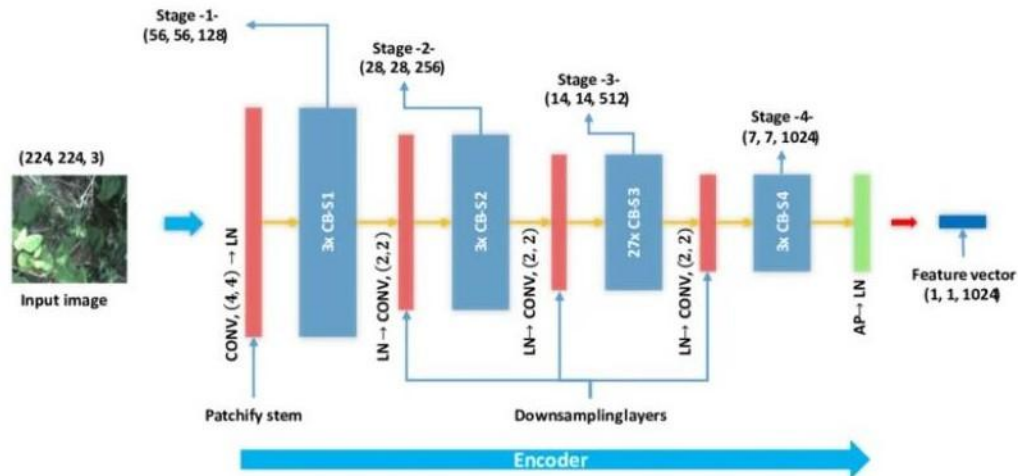


Gambar 2.2 Grafik Benchmark CNN dan ViT
Sumber: (Liu et al., 2022)

Berdasarkan Gambar 2.2, terlihat jelas alasan mengapa arsitektur ini sangat direkomendasikan. Grafik tersebut membandingkan tingkat akurasi (sumbu vertikal) dan beban komputasi atau *GFLOPs* (direpresentasikan oleh diameter gelembung) antara keluarga CNN (gelembung biru) dan keluarga *Vision Transformer* (gelembung hijau). Terlihat bahwa ConvNeXt berhasil menduduki posisi puncak dengan akurasi tertinggi, khususnya pada skenario pelatihan dataset skala masif (ImageNet-22K), mengalahkan arsitektur *Transformer* mutakhir seperti Swin Transformer dan ViT. Pencapaian akurasi maksimal ini berhasil diraih dengan tetap mempertahankan efisiensi beban komputasi yang setara.

ConvNeXt mengadopsi beberapa perubahan struktur kunci, di antaranya penggunaan filter *depthwise convolution* berukuran 7×7 , struktur *inverted bottleneck*, fungsi aktivasi GELU menggantikan ReLU, serta penggunaan *LayerNorm*

menggantikan *BatchNorm*. Secara rinci, alur pemrosesan hierarkis citra pada arsitektur ConvNeXt dari tahap *input* hingga klasifikasi akhir diilustrasikan pada Gambar 2.3 berikut:



Gambar 2. 3 Arsitektur ConvNeXt
Sumber: (Benchallal et al., 2024)

Berdasarkan Gambar 2.3, alur kerja pemrosesan citra pada model ConvNeXt dapat diuraikan ke dalam tiga tahapan utama:

1. Tahap Masukan dan *Stem Layer*: Citra masukan berukuran 224 x 224 piksel dengan 3 saluran warna (RGB) diproses pertama kali oleh lapisan awal. Lapisan konvolusi awal ini mereduksi resolusi spasial gambar menjadi 56 x 56 piksel dan mengekstraksi 128 peta fitur, yang kemudian dinormalisasi menggunakan *Layer Norm* agar distribusi data lebih stabil.
2. Tahap *ConvNeXt Blocks* (Ekstraksi Fitur Hierarkis): Tahapan ini merupakan inti ekstraksi pola yang dibagi menjadi empat tahapan (*stages*) dengan blok konvolusi yang diulang secara spesifik pada varian *Base* (sebanyak 3, 3, 27, dan 3 kali). Di antara tahapan tersebut, terdapat lapisan *Down Sample* yang bertugas memperkecil resolusi spasial gambar secara bertahap (dari 56 x 56, 28 x 28, 14 x 14, hingga 7 x 7) sekaligus memperdalam dimensi fitur (dari 128, 256, 512, hingga 1024). Pendekatan mengerucut ini memungkinkan model menangkap detail beresolusi tinggi di awal (seperti tekstur lesi penyakit) dan merangkum konteks visual secara keseluruhan di akhir.

3. Tahap Klasifikasi Akhir (*Classification Head*): Fitur tingkat tinggi yang terkumpul dengan dimensi $7 \times 7 \times 1024$ diratakan menggunakan lapisan *Global Average Pooling* menjadi matriks tunggal ($1 \times 1 \times 1024$). Vektor fitur ini kembali dinormalisasi oleh *Layer Norm*, dilewatkan pada lapisan saraf linear (*Linear*), dan diakhiri dengan fungsi *Softmax* untuk memetakan representasi tersebut menjadi nilai probabilitas kelas keluaran akhir.

2.2.8 Transfer Learning

Transfer learning merupakan teknik dalam *deep learning* yang memanfaatkan model yang telah dilatih sebelumnya pada dataset besar untuk diterapkan pada tugas baru dengan skala dataset kecil atau terbatas (Böhle et al., 2021). Teknik ini sangat krusial digunakan karena model *deep learning* tradisional membutuhkan jumlah data dan waktu pelatihan yang sangat besar agar dapat mempelajari representasi fitur dengan baik. Dalam kasus klasifikasi penyakit daun sirih di mana jumlah citra lapangan yang tersedia relatif terbatas, pendekatan ini membantu mempercepat konvergensi waktu komputasi sekaligus meminimalkan risiko *overfitting*.

Pada penelitian ini, proses *transfer learning* (*fine-tuning*) diimplementasikan melalui dua pendekatan sumber bobot (*pre-trained weights*) yang sangat mutakhir. Pada model ConvNeXt, representasi fitur lokal diekstraksi menggunakan bobot dasar yang telah dilatih pada dataset berskala masif ImageNet-22k, sebuah repositori yang berisi belasan juta citra dengan 22 ribu kelas, sehingga model memiliki kepekaan tinggi terhadap variasi tekstur dan pola warna. Sementara itu, pada arsitektur DINOv2, bobot awal yang digunakan bersumber dari paradigma *Self-Supervised Learning*, di mana model telah belajar mengenali relasi fitur spasial secara mandiri dari kumpulan citra raksasa tanpa melibatkan pelabelan manusia. Penggabungan *pre-trained weights* dari kedua domain sumber ini memastikan model dapat beradaptasi secara optimal saat dihadapkan pada target domain, yaitu citra daun sirih lokal.

2.2.9 Transformer dan Self-Attention

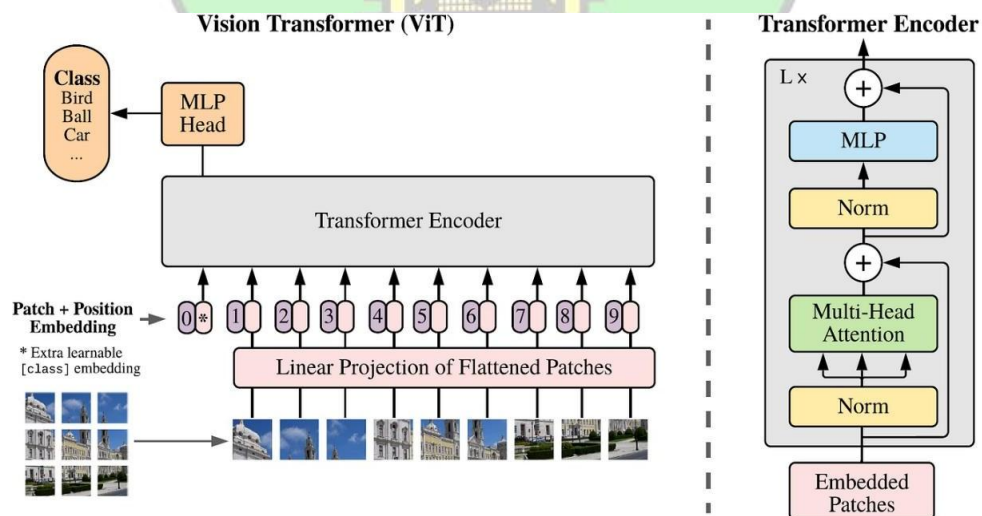
Transformer adalah arsitektur jaringan saraf yang revolusioner, awalnya dirancang untuk tugas pemrosesan bahasa alami (NLP) namun kemudian diadaptasi

sukses untuk visi komputer (seperti pada Vision Transformer). Berbeda dengan arsitektur rekuren (RNN) atau konvolusional (CNN) tradisional, Transformer sepenuhnya mengandalkan mekanisme Self-Attention untuk memproses data sekuensial.

Menurut penelitian (Wahid et al., 2024), kekuatan Transformer terletak pada kemampuannya untuk menangkap dependensi jarak jauh antar elemen dalam sebuah sekuens secara efisien, tanpa harus memproses data secara berurutan. Hal ini menjadikannya sangat efektif dalam memahami konteks global dan hubungan kompleks antar-bagian citra.

2.2.10 Vision Transformer

Vision Transformer (ViT) adalah arsitektur *Deep learning* yang mengadaptasi model Transformer yang awalnya sangat sukses di bidang teks (NLP) untuk digunakan dalam pengolahan citra digital. Berbeda dengan CNN yang menggunakan operasi konvolusi untuk memindai gambar secara lokal, ViT memperlakukan sebuah gambar sebagai urutan potongan (sequence of patches) (Y. Wang et al., 2025). Keunggulan utama ViT adalah kemampuannya untuk menangkap hubungan antar bagian gambar secara menyeluruh atau fitur global, sehingga model dapat memahami konteks spasial yang luas dari sebuah citra, seperti yang ditunjukkan pada gambar 2.4.



Gambar 2.4 Arsitektur Vision Transformer
Sumber: (Wong, 2025)

Berdasarkan arsitektur standarnya, proses kerja ViT dalam mengenali pola (seperti penyakit pada daun sirih) terdiri dari tahapan berikut :

1. *Image Patching* (Potongan gambar)

Pembagian potongan gambar (misalnya dengan ukuran 224x224) menjadi potongan kotak kecil yang disebut *patches* (misalnya ukuran 16x16). Tahapan ini mengubah satu unit citra menjadi 196 urutan potongan gambar yang dapat diolah secara matematis oleh arsitektur Transformer.

2. *Patch Embedding*

Setiap patch yang masih berupa data piksel dua dimensi (2D) diratakan (*flattened*) menjadi vektor satu dimensi (1D). Vektor ini kemudian diproyeksikan menggunakan matriks bobot yang dapat dilatih (*linear projection*) ke dalam ruang dimensi yang lebih tinggi. Hasilnya adalah kumpulan *patch embeddings* yang merepresentasikan ciri visual dari tiap kepingan gambar.

3. Penambahan *Classification Token (CLS)*

Penambahan satu token khusus yang disebut *Classification Token (CLS)* di awal urutan patch embedding. Token ini bertindak sebagai "rangkuman" digital. Selama proses pelatihan, token CLS akan berinteraksi dengan seluruh patch gambar lainnya melalui mekanisme perhatian, sehingga pada lapisan akhir, token ini membawa informasi representasi global dari seluruh citra untuk kebutuhan klasifikasi.

4. *Positional Embedding*

Karena model *Transformer* tidak mengenali urutan lokasi secara alami, maka ditambahkan informasi posisi pada setiap patch. Hal ini memastikan model mengetahui letak setiap kepingan di dalam gambar utuh (seperti mana bagian tepi atau tengah objek)

5. *Transformer Encoder*

Urutan token tersebut dimasukkan ke dalam blok Transformer Encoder. Di dalamnya terdapat mekanisme *Multi-Head Self-Attention* yang

memungkinkan setiap bagian gambar "memperhatikan" bagian lainnya secara simultan. Hal ini sangat membantu dalam menangkap hubungan jarak jauh (fitur global), seperti bagaimana pola warna di ujung daun berkaitan dengan tekstur di pangkal daun.

6. *Multi-Layer Perceptron (MLP) Head*

Pada tahap akhir, representasi fitur yang telah diperkaya dari token CLS dikeluarkan dari *Encoder* dan diteruskan ke *MLP Head*. Bagian ini terdiri dari lapisan dan fungsi aktivasi yang memetakan fitur global tersebut menjadi probabilitas kelas penyakit, yang memungkinkan sistem untuk memberikan hasil akhir apakah daun sirih tersebut sehat atau terinfeksi patogen.

Pada Tabel 2.2 Model Vision transformer dibagi dalam beberapa varian berdasarkan jumlah parameter, kedalaman lapisan (*depth*), dan dimensi vektor fitur (hidden size) yaitu ViT-Base, ViT-Large, dan ViT-Huge (Dosovitskiy et al., 2021).

Tabel 2.2 *Varian Model Vision Transformer*

Model	Layers	Hidden size D	MPL size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Perbedaan utama antara ViT-Base, ViT-Large, dan ViT-Huge terletak pada jumlah lapisan (layers), dimensi vektor fitur (hidden size), jumlah attention heads, serta total parameter model. Semakin besar varian ViT yang digunakan, semakin dalam dan kompleks arsitekturnya, sehingga kemampuan model dalam menangkap representasi fitur global juga meningkat, namun diikuti oleh kebutuhan komputasi dan data pelatihan yang lebih besar. Oleh karena itu, pemilihan varian ViT harus disesuaikan dengan kompleksitas tugas, ukuran dataset, dan keterbatasan sumber daya komputasi,

di mana ViT-Base umumnya lebih efisien untuk dataset menengah, sedangkan ViT-Large dan ViT-Huge cocok untuk skala data besar dengan kebutuhan representasi fitur yang lebih kaya (Dosovitskiy et al., 2021).

2.2.11 DINOv2 (Distillation with NO labels)

DINOv2 adalah arsitektur *Vision Transformer* mutakhir yang dikembangkan oleh Meta AI, dibangun di atas paradigma pembelajaran mandiri (*Self-Supervised Learning*) yang tidak memerlukan anotasi label manusia selama fase *pre-training* (Oquab et al., 2024). Berbeda dengan *Vision Transformer* konvensional yang membutuhkan jutaan citra berlabel untuk dilatih, DINOv2 mempelajari representasi visual dengan mencocokkan fitur pada tingkat citra (*image-level*) maupun tingkat kepingan gambar (*patch-level*) melalui mekanisme *self-distillation*. Pendekatan ini memungkinkan model untuk belajar mengenali pola dari konsistensi hasil prediksinya sendiri terhadap variasi (augmentasi) dari citra yang sama, tanpa memerlukan label eksternal dari manusia.

Secara arsitektural, proses kerja dasar DINOv2 varian *base* tersusun atas 12 lapisan *Transformer encoder* dengan 12 *attention head* per lapisan. Saat menerima masukan gambar, proses dimulai dengan memecah citra input menjadi kepingan-kepingan kecil berukuran 14×14 piksel (*patch embedding*). Potongan-potongan ini kemudian diberi informasi posisi (*positional embedding*) sebelum diproses melalui mekanisme *Multi-Head Self-Attention* secara berlapis. Keluaran akhir dari proses ini direpresentasikan dalam bentuk *Class Token* (CLS token) yaitu sebuah vektor fitur tunggal yang merangkum keseluruhan konteks visual citra secara global.

Sifat *self-supervised* pada DINOv2 pada dasarnya hanya berlaku pada tahap *pre-training* yang telah dilakukan secara eksklusif oleh Meta AI menggunakan dataset skala masif (LVD-142M). Bobot hasil *pre-training* ini bersifat dapat digunakan kembali (*reusable*) sebagai titik awal (*pre-trained weights*).

Pada penerapan dalam penelitian ini, bobot dasar facebook/dinov2-base tersebut diadaptasi menggunakan metode pembelajaran terarah (*supervised fine-*

tuning). Proses ini dilakukan dengan menambahkan sebuah *classification head* baru, yaitu lapisan linear di atas jaringan DINOv2 yang bertugas memetakan 768 dimensi fitur dari token CLS menjadi 3 kelas keluaran klasifikasi. Lapisan baru ini, beserta penyesuaian bobot arsitekturnya, dilatih ulang secara spesifik menggunakan data citra daun sirih yang telah diberi label kelas (*Healthy*, *Bacterial*, dan *Dried*). Selama proses pelatihan, jaringan ini mengevaluasi tebakannya menggunakan fungsi kerugian *cross-entropy loss* dan memperbarui bobotnya melalui proses *backpropagation* agar model mampu mendiagnosis kondisi daun sirih secara akurat.

2.2.12 Evaluasi Performa

Untuk mengukur performa dan efektivitas model *Deep learning* dalam mengklasifikasikan citra, diperlukan parameter ukur yang objektif. Performa model dinilai berdasarkan perbandingan antara hasil prediksi sistem dengan label sebenarnya (*ground truth*). Dasar dari perhitungan metrik evaluasi ini adalah *Confusion Matrix* yang melibatkan empat parameter utama:

- a) *True Positive (TP)*: Jumlah data positif yang diprediksi benar sebagai positif.
- b) *True Negative (TN)*: Jumlah data negatif yang diprediksi benar sebagai negatif.
- c) *False Positive (FP)*: Jumlah data negatif yang diprediksi salah sebagai positif (*Peringatan Palsu*).
- d) *False Negative (FN)*: Jumlah data positif yang diprediksi salah sebagai negatif.

Adapun metrik evaluasi yang digunakan dalam penelitian ini meliputi:

1. Akurasi (*Accuracy*)

Akurasi mengukur kemampuan model dalam mengklasifikasikan seluruh kondisi daun (Penyakit daun), berikut formulasi pengukur :

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad 3.2.1$$

2. Presisi (Precision)

Presisi menunjukkan seberapa akurat model saat mendeteksi penyakit. Metrik ini meminimalkan kesalahan model dalam memberikan “Peringatan Palsu”.

Perhitungan presisi dapat dilakukan dengan persamaan berikut :

$$Precision = \frac{TP}{TP + FP} \quad 3.2.2$$

3. Sensitivitas (Recall)

Recall mengukur kemampuan model dalam mendeteksi seluruh sampel penyakit yang ada. Metrik ini sangat penting agar tidak ada daun yang sakit tetapi dianggap sehat oleh sistem. Berikut formulasinya :

$$Recall = \frac{TP}{TP + FN} \quad 3.2.3$$

4. F1-Score

F1-Score memberikan nilai rata-rata yang menyeimbangkan antara Presisi dan Recall. Indikator ini dinyatakan dalam hubungan matematis sebagai berikut:

$$F1 = 2X \frac{Recall \times Precision}{Recall + Precision} \quad 3.2.4$$

Keterangan : *TP* (True Positive) *TN* (True Negative) *FP* (False Positive) *FN*(False Negative)

BAB III

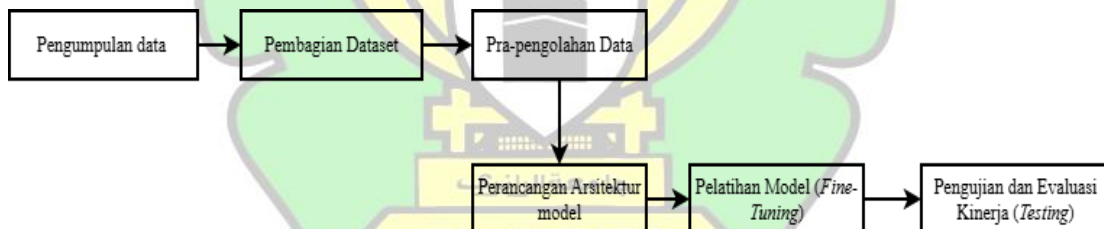
METODOLOGI PENELITIAN

3.1 Jenis Penelitian

Penelitian ini merupakan penelitian eksperimental dengan pendekatan kuantitatif yang berfokus pada studi komparatif performa model *deep learning*. Secara spesifik, penelitian ini menggunakan metode perancangan sistem (*System Design*) untuk mengimplementasikan dan membandingkan dua arsitektur mutakhir secara independen, yaitu ConvNeXt (representasi arsitektur CNN modern) dan DINOv2 (representasi *Self-Supervised Vision Transformer*), dalam tugas klasifikasi citra digital penyakit daun sirih.

Analisis data dilakukan dengan membandingkan hasil prediksi kedua model terhadap label aktual menggunakan metrik evaluasi objektif guna mengukur tingkat efektivitas masing-masing arsitektur dalam mengenali pola penyakit secara akurat.

3.2 Tahapan Penelitian



Gambar 3.1 Alur Penelitian

Tahapan penelitian ini dirancang secara sistematis dengan mengikuti metodologi pengembangan sistem *deep learning* yang komprehensif, dimulai dari fase pengumpulan data, pra-pengolahan citra otomatis, perancangan arsitektur, pelatihan model, hingga tahap evaluasi akhir dan interpretasi visual. Pendekatan ini bertujuan untuk menciptakan alur kerja yang terintegrasi, di mana setiap fasenya memberikan kontribusi krusial dalam mengevaluasi serta membandingkan efektivitas ekstraksi fitur

spasial lokal pada arsitektur CNN modern (ConvNeXt) dan representasi fitur global pada *Vision Transformer* (DINOv2) guna mengidentifikasi penyakit pada daun sirih secara akurat.

3.2.1 Pengumpulan Data

Dalam pembangunan model yang lebih terstruktur dan mampu dalam menjalankan penelitian yang memungkinkan sesuai dengan standar variasi standar lapangan, penelitian ini akan menerapkan pendekatan penggabungan dataset (kombinasi dataset) yaitu data sekunder dan primer dengan total 2309 citra.

a. Data Sekunder

Data sekunder dalam penelitian ini diperoleh untuk meningkatkan kemampuan generalisasi model dalam mengenali berbagai pola visual penyakit daun sirih dari berbagai kondisi lingkungan yang berbeda. Data sekunder dalam penelitian ini diperoleh dari pemanfaatan dataset publik dari repositori digital terkemuka seperti Kaggle guna memperluas variasi sampel citra penyakit tanaman.

Dalam pemanfaatan data sekunder koleksi citra yang terkumpul diklasifikasikan ke dalam tiga kategori utama untuk memenuhi target klasifikasi model. Kelas-kelas tersebut meliputi kondisi " *Healthy Leaf* ", *Bacterial Leaf Disease*, dan *Dried Leaf*". Secara keseluruhan, pemanfaatan dataset publik dari repositori ini menyediakan total 1.934 citra guna memperluas variasi sampel penyakit tanaman pada kondisi ideal.

b. Data Primer

Diperoleh melalui observasi dan pengambilan citra langsung di perumahan penduduk Darussalam menggunakan kamera smartphone Realme 6 Pro. Penggunaan resolusi kamera tinggi dari perangkat ini memungkinkan penangkapan detail tekstur penyakit daun sirih secara presisi di bawah kondisi pencahayaan alami lingkungan sekitar. Total data primer yang dikumpulkan adalah 375 citra. Seluruh kumpulan dataset ini dikategorikan ke dalam tiga kelas utama, yaitu *Healthy Leaf* , *Bacterial Leaf Disease*, dan *Dried Leaf* .

3.2.2 Persiapan dan Pengolahan Data

Langkah ini mencakup penyeselarasan karakteristik data dari berbagai sumber akuisisi untuk menjamin kualitas dan integritas masukan selama fase pelatihan model. Proses persiapan data pada tahap ini mencakup, antara lain :

1. *Splitting* (Pembagian Data).

Guna memastikan evaluasi model berjalan secara objektif dan mampu merepresentasikan generalisasi yang sebenarnya, dataset dipartisi ke dalam tiga himpunan: Data Latih (*Training Set*), Data Validasi (*Validation Set*), dan Data Uji (*Test Set*). Pemisahan ini dieksekusi dalam dua tahap utama, diawali dengan pengisolasian citra ke dalam folder *test* secara fisik agar himpunan tersebut sama sekali tidak terpapar selama masa pelatihan dan murni berfungsi sebagai instrumen pengujian akhir (*blind test*).

Selanjutnya, himpunan citra pada folder *train* dipartisi kembali dengan rasio 85:15 menjadi Data Latih dan Data Validasi menggunakan fungsi *train_test_split* dari pustaka *Hugging Face*. Proses partisi tingkat kedua ini mengimplementasikan teknik stratifikasi (*stratified split*) yang dikendalikan oleh nilai *seed* tetap, sehingga keseimbangan proporsi jumlah citra pada masing-masing kelas dapat terus terjaga dan keseluruhan eksperimen dapat direproduksi ulang secara konsisten.

2. *Preprocessing* (Pra-pengolahan).

Guna menstandarisasi format masukan secara konsisten, tahapan pra-pengolahan citra dasar dalam penelitian ini dieksekusi dan diotomatisasi secara komprehensif menggunakan modul *AutoImageProcessor* yang terintegrasi di dalam ekosistem *Hugging Face*. Modul cerdas ini bertugas menyelaraskan setiap matriks citra mentah agar secara spesifik memenuhi standar input dari arsitektur *pre-trained facebook/convnext-base-224-22k* dan *facebook/dinov2-base*.

Proses transformasi yang berjalan pada modul ini mencakup beberapa penyesuaian matematis tingkat matriks: pertama, penyesuaian resolusi dan

pemotongan (*resizing and cropping*) citra secara proporsional menjadi matriks standar berukuran 224x224 piksel; kedua, konversi saluran warna ke format *Red-Green-Blue* (RGB) 3-dimensi yang ditranslasikan ke dalam bentuk tipe data Tensor PyTorch; dan ketiga, penerapan normalisasi distribusi nilai piksel (*pixel normalization*). Operasi normalisasi ini menyesuaikan nilai rata-rata (*mean*) dan standar deviasi (*standard deviation*) citra berdasarkan metrik pemusatan dataset ImageNet, sebuah tahapan matematis yang esensial untuk menstabilkan kalkulasi *gradient descent* dan mempercepat titik konvergensi (pembelajaran optimal) model selama iterasi *training* berlangsung.

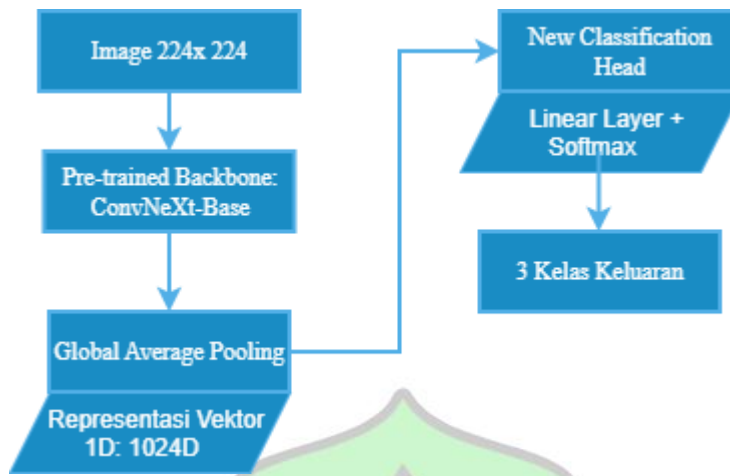
3. *Data Augmentation (Augmentasi Data)*

Selain pra-pengolahan dasar, penelitian ini juga mengimplementasikan teknik augmentasi data dinamis menggunakan pustaka *torchvision.transforms*. Transformasi ini secara eksklusif hanya diterapkan pada himpunan Data Latih untuk meningkatkan ketangguhan (*robustness*) model terhadap variasi kondisi visual di lapangan terbuka. Parameter manipulasi yang diaplikasikan meliputi *Random Resized Crop* (pemotongan dan penyesuaian ukuran acak), *Random Horizontal Flip* (pembalikan gambar secara *horizontal*), serta *Random Rotation* hingga 15 derajat. Lebih lanjut, fitur *Color Jitter* digunakan untuk menyimulasikan perubahan kondisi pencahayaan alami (kecerahan, kontras, saturasi, dan rona warna), serta efek *Gaussian Blur* diterapkan secara acak guna melatih model agar kebal terhadap hasil tangkapan kamera seluler yang kurang fokus. Himpunan Data Validasi dan Data Uji sama sekali tidak melalui proses augmentasi ini agar evaluasi metrik kinerja tetap objektif dan murni.

3.2.3 Perancangan Arsitektur Model

Penelitian ini merancang dua jalur klasifikasi independen guna membandingkan efektivitas ekstraksi fitur lokal dan global:

1. Jalur ConvNeXt



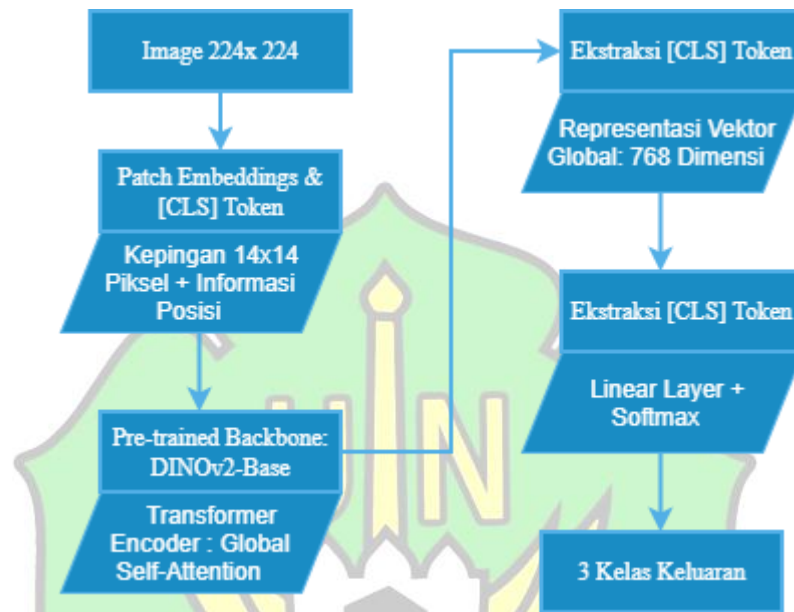
Gambar 3.2 Arsitektur Perancangan ConvNeXt pada Daun Sirih

Berdasarkan gambar 3.2, berikut Proses klasifikasi menggunakan arsitektur ConvNeXt.

- Backbone* (Tulang Punggung Ekstraksi Fitur): Bagian ini adalah arsitektur bawaan ConvNeXt-Base yang bobotnya didapatkan dari hasil pelatihan dataset raksasa ImageNet-22k. Tugas utama bagian ini (dari *Stem* hingga *Stage 4*) murni hanya untuk memindai dan mengekstraksi fitur visual, seperti tekstur permukaan daun, lekukan, gradasi warna kuning pada daun kering, hingga bercak hitam mikro pada daun bakteri.
- Pemotongan (Modifikasi Arsitektur): Model asli ConvNeXt dari *Hugging Face* secara *default* akan mengeluarkan puluhan ribu kelas tebakan bawaan ImageNet. Pada tahap ini, arsitektur dimodifikasi secara paksa dengan membuang ujung lapisan klasifikasi (*classifier.weight* dan *classifier.bias*) tersebut.
- New Classification Head* (Kepala Klasifikasi Baru): Setelah lapisan lama dibuang, model dipasangkan sebuah lapisan saraf *linear* baru. Lapisan ini bertindak sebagai jembatan yang menerjemahkan

kumpulan vektor fitur spasial yang sudah diekstraksi oleh *Backbone*, dan memetakannya secara presisi ke dalam ruang probabilitas untuk 3 kelas target spesifik penelitian Anda (Sehat, Bakteri, Kering).

2. Jalur DINOv2



Gambar 3.3 Arsitektur Perancangan DINOv2 pada Daun Sirih

Berdasarkan gambar 3.3, berikut Proses klasifikasi menggunakan arsitektur DINOv2.

- a. Pemecahan dan Representasi Awal (*Patching & Embedding*): Berbeda dengan CNN yang memindai gambar utuh, arsitektur ini pertama-tama akan "memotong" citra daun sirih menjadi kepingan-kepingan kecil (*patches*) berukuran 14x14 piksel. Setiap kepingan kemudian diubah menjadi angka (*Patch Embeddings*) dan disisipkan informasinya (*Positional Embeddings*) agar model tahu susunan asli daun tersebut. Di tahap ini pula, sebuah token spesial bernama [CLS] Token (*Class Token*) dimasukkan di urutan paling awal.

- b. Mekanisme Perhatian Global (*Transformer Encoder*): Potongan-potongan tersebut kemudian masuk ke dalam "otak" utama DINOv2 (facebook/dinov2-base) yang sebelumnya sudah belajar mandiri melalui paradigma *Self-Supervised Learning*. Di dalam lapisan ini, model menggunakan mekanisme *Global Self-Attention*, di mana setiap kepingan gambar daun dipaksa untuk mencari tahu hubungannya dengan kepingan lain secara menyeluruh.
- c. Ekstraksi Fitur Global ([CLS] Token): Jika ConvNeXt menggunakan *Global Average Pooling* di akhir untuk merangkum fitur, DINOv2 menggunakan cara yang berbeda. Di pintu keluar *Encoder*, model hanya akan mengambil vektor dari [CLS] Token tersebut. Karena [CLS] Token ini sudah berinteraksi dengan seluruh kepingan daun di tahap sebelumnya, maka ia kini berisi rangkuman fitur global dari gambar daun tersebut dalam bentuk 768 dimensi.
- d. Kepala Klasifikasi Baru (*New Classification Head*): Lapisan proyektor asli dari DINOv2 dibuang, dan sebagai gantinya, vektor [CLS] berdimensi 768 tadi disalurkan ke sebuah lapisan klasifikasi linear yang baru (*Classification Head*). Lapisan baru ini khusus dilatih (*fine-tuning*) agar bisa mengerucutkan 768 informasi fitur tersebut menjadi probabilitas untuk menebak 3 kelas secara presisi: Daun Sehat, Bakteri, atau Kering.

3.2.4 Pelatihan Model

Proses pelatihan atau penyesuaian bobot (*fine-tuning*) dikelola secara efisien menggunakan antarmuka Trainer API dari ekosistem *Hugging Face*. Efektivitas pembelajaran diatur melalui konfigurasi *hyperparameter* berikut:

1. Learning Rate: Ditetapkan sebesar $2e-5$ untuk menjamin stabilitas pembaruan bobot dan mencegah *overfitting* pada model *pre-trained*.
2. Batch Size: Ditetapkan sebesar 16 guna menyeimbangkan akurasi estimasi gradien dengan kapasitas memori komputasi.

3. Epoch: Total 15 epoch ditentukan sebagai batas iterasi, karena konvergensi optimal telah tercapai pada rentang iterasi tersebut.
4. Weight Decay: 0.05 sebagai teknik regularisasi untuk mencegah terjadinya *overfitting*.
5. Optimizer: Menggunakan AdamW untuk percepatan konvergensi gradien.
6. Checkpoints: Menerapkan fungsi `load_best_model_at_end=True` yang secara otomatis menyimpan pembobotan terbaik berdasarkan metrik akurasi tertinggi pada fase validasi.

3.3 Bahan dan Alat Penelitian

Pada penelitian ini menyediakan perangkat keras dan perangkat lunak sebagai alat dan bahan penelitian, adapun sebagai berikut :

1. *Hardware* (Perangkat Keras)
 - a. Lenovo V14-IIL (Model 82C4), untuk *Processor* INTEL CORE i5-1035G1 (Generasi ke-10 / Ice Lake, 4 Core 8 Thread), dengan RAM: 12 GB dan Storage 256 GB SSD (M.2 NVMe).
 - b. Camera Realme 6 Pro sebagai alat untuk pengambilan gambar.
 - c. Proses pengembangan ataupun training model dijalankan pada salah satu layanan google yaitu *Google Colab* (*Google Colaboratory*)
2. *Software* (Perangkat Lunak)
 - a. Bahasa Pemrograman: *Python*, digunakan sebagai bahasa utama untuk seluruh proses pengolahan data dan pembangunan model.
 - b. Lingkungan Pengembangan: *Google Colab* atau *Jupyter Notebook*, sebagai platform eksekusi kode berbasis cloud yang menyediakan akselerasi GPU.
 - c. *Hugging Face* Transformers, sebagai pustaka utama penyedia arsitektur *state-of-the-art (SOTA)* untuk memuat bobot model *pre-trained* (ConvNeXt dan DINOv2) sekaligus memfasilitasi modul *AutoImageProcessor* untuk pra-pengolahan citra dan *Trainer API* untuk eksekusi *fine-tuning*."

- d. *PyTorch*, bertindak sebagai mesin penggerak di balik library.
- e. Library Visi Komputer: *OpenCV* dan *torchvision*, untuk proses pra-pengolahan citra seperti *resizing* dan normalisasi.

Library Analisis Data: *NumPy* dan *Pandas*, untuk manajemen dataset serta *Matplotlib* untuk visualisasi hasil evaluasi.



BAB IV

HASIL DAN PEMBAHASAN

4.1 Skenario Pengujian

Tahap ini menguraikan skenario yang telah dirancang sebelum model mulai dieksekusi, yang meliputi struktur pembagian dataset serta penentuan parameter konfigurasi (*hyperparameter*) yang ditanamkan pada saat proses pelatihan.

4.1.1 Persiapan dan Pembagian Dataset

Pada Penelitian ini agar pengujian berjalan secara objektif dan mampu merepresentasikan kinerja model pada kondisi klasifikasi yang sebenarnya, dataset yang telah diakuisisi harus melalui tahap persiapan dan pembagian secara terstruktur. Dalam penelitian ini, seluruh citra dikategorikan ke dalam tiga kelas utama, yaitu *Healthy Leaf*, *Bacterial Leaf Disease*, dan *Dried Leaf*.

Sebagaimana disebutkan pada Bab III, keseluruhan arsip dataset yang diakuisisi berjumlah 2.309 berkas. Namun, pada tahap pemuatan data ke dalam sistem, dilakukan proses penyaringan (*filtering*) berdasarkan format citra (*.jpg*, *.jpeg*, *.png*) untuk memastikan hanya berkas citra yang valid yang diikutsertakan dalam proses pemodelan. Proses penyaringan ini menyisihkan 2 berkas non-citra yang turut ter-ekstrak dari arsip dataset, sehingga diperoleh total 2.307 citra valid yang benar-benar digunakan dalam penelitian ini.

Total keseluruhan citra yang valid dan diproses untuk pemodelan berjumlah 2.307 citra. Dari total tersebut, sistem secara spesifik mendeteksi 2.100 citra pada direktori latih dan 207 citra terpisah pada direktori uji.

Guna mengoptimalkan proses pemantauan performa model selama fase *fine-tuning* berlangsung, kumpulan 2.100 citra dasar tersebut dipartisi kembali secara otomatis di dalam sistem dengan rasio 85:15. Proses ini menghasilkan alokasi sebesar 1.785 citra sebagai Data Latih (*Training Set*) dan 315 citra sebagai Data Validasi (*Validation Set*). Pembagian ini dieksekusi menggunakan fungsi stratifikasi (*stratified*

split) yang menjamin bahwa proporsi jumlah gambar dari ketiga kelas daun tetap seimbang pada masing-masing himpunan data.

Sementara itu, sisa 207 citra lainnya benar-benar diisolasi dari awal dan digunakan secara eksklusif sebagai Data Uji (*Test Set*). Pendekatan isolasi data ini sangat krusial untuk memastikan bahwa evaluasi akhir (*blind test*) benar-benar dilakukan menggunakan citra yang belum pernah dilihat oleh model selama proses pelatihan.

Sebagai tahap akhir dari persiapan ini, seluruh citra dalam dataset diproses secara dinamis menggunakan modul *AutoImageProcessor*. Modul ini bertugas menyeragamkan dimensi gambar dengan mengubah resolusinya menjadi matriks standar berukuran 224x224 piksel serta mengaplikasikan normalisasi tensor agar data siap diolah oleh arsitektur *ConvNeXt* maupun *DINOv2*.

Penting untuk digaris bawahi bahwa seluruh teknik augmentasi data dan manipulasi visual buatan tersebut secara eksklusif hanya diberlakukan pada fase pelatihan (Data Latih). Himpunan Data Validasi dan Data Uji dipertahankan dalam kondisi murni (tanpa distorsi atau augmentasi buatan) guna menjamin objektivitas evaluasi metrik kinerja model pada tahap akhir pengujian. Rincian distribusi dataset ini disajikan secara lengkap pada Tabel 4.1.

Tabel 4.1 *Rincian Pembagian Dataset Daun Sirih*

Jenis Dataset	Keterangan / Rincian Kelas	Jumlah Citra
Data latih	85% dari alokasi data latih dasar. Digunakan untuk membangun pola model.	1.785
Data Validasi	15% dari alokasi data latih dasar. Digunakan untuk evaluasi saat proses pelatihan berjalan.	315
Data Uji	<i>Healthy Leaf</i> (69), <i>Bacterial Leaf Disease</i> (67), <i>Dried Leaf</i> (71). Digunakan untuk pengujian buta (<i>blind test</i>).	207

4.1.2 Hyperparameter Model

Agar perbandingan antara model Convolutional Neural Network (ConvNeXt) dan Vision Transformer (DINOv2) berjalan secara adil, kedua arsitektur tersebut dilatih menggunakan konfigurasi hyperparameter yang seragam melalui modul *Trainer*. Rincian konfigurasi yang digunakan pada proses pelatihan ditunjukkan pada Tabel 4.2.

Tabel 4.2 Konfigurasi Hyperparameter Pelatihan

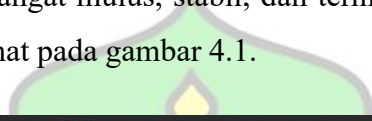
Hyperparameter	Nilai	Justifikasi
epoch	15	Memberikan jumlah iterasi yang cukup bagi model untuk mencapai titik konvergensi (pembelajaran maksimal) tanpa memicu <i>overfitting</i> .
Learning Rate	2e-5 (0.00002)	Nilai yang kecil sangat optimal untuk proses <i>fine-tuning</i> model <i>pre-trained</i> agar tidak merusak bobot (weights) dasar yang sudah ada.
Batch Size	16	Menyeimbangkan kestabilan pembaruan gradien sekaligus mencegah kelebihan beban memori (<i>Out of Memory</i>) pada GPU <i>Google Colab</i> .
Weight Decay	0.05	Berfungsi sebagai teknik regularisasi untuk mengontrol kompleksitas model, sehingga model dapat menggeneralisasi data baru dengan lebih baik
Optimizer	AdamW	Mempercepat konvergensi gradien secara adaptif dengan regularisasi bobot yang efisien.

4.2 Hasil pengujian

Tahap ini memaparkan hasil dari proses pelatihan (*fine-tuning*) yang telah dijalankan selama 15 epoch untuk masing-masing model, dilanjutkan dengan evaluasi kinerja (pengujian) menggunakan 207 citra data uji eksklusif.

4.2.1 Hasil Pelatihan ConvNeXt

Proses pelatihan model ConvNeXt dievaluasi selama 15 epoch untuk mengamati pergerakan nilai kesalahan (*loss*) dan metrik akurasi pada data latih serta data validasi. Berdasarkan hasil perekaman log pelatihan, model CNN modern ini menunjukkan proses konvergensi yang sangat mulus, stabil, dan terhindar dari indikasi *overfitting*. Proses pelatihan dapat dilihat pada gambar 4.1.



Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.041498	0.104143	0.965079	0.965242	0.965940	0.965079
2	0.028717	0.115030	0.971429	0.971415	0.972516	0.971429
3	0.079041	0.071282	0.984127	0.984125	0.984505	0.984127
4	0.002394	0.065902	0.980952	0.980967	0.981186	0.980952
5	0.000246	0.062770	0.980952	0.980948	0.981556	0.980952
6	0.000235	0.044823	0.987302	0.987317	0.987362	0.987302
7	0.000168	0.075871	0.980952	0.980967	0.981186	0.980952
8	0.013013	0.050024	0.984127	0.984142	0.984246	0.984127
9	0.000117	0.045471	0.987302	0.987301	0.987508	0.987302
10	0.025794	0.038392	0.987302	0.987301	0.987508	0.987302
11	0.000067	0.073424	0.977778	0.977769	0.978660	0.977778
12	0.000066	0.092556	0.974603	0.974587	0.975814	0.974603
13	0.003463	0.042532	0.990476	0.990476	0.990566	0.990476
14	0.000143	0.040784	0.987302	0.987317	0.987362	0.987302
15	0.030054	0.039824	0.987302	0.987317	0.987362	0.987302

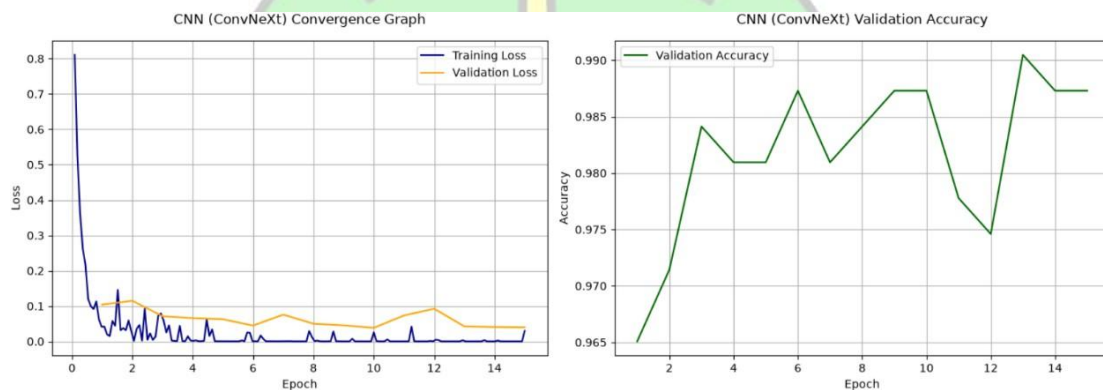
Gambar 4. 1 Log metrik pelatihan per-epoch pada ConvNeXt

Berdasarkan gambar 4.1 Pergerakan nilai *loss* dan akurasi selama proses pelatihan dapat dirincikan sebagai berikut:

1. Kondisi Awal Pelatihan: Pada epoch pertama, model langsung menunjukkan adaptasi yang baik dengan mencatatkan nilai *training loss* sebesar 0.0414 dan *validation loss* 0.1041, disertai tingkat akurasi validasi 96.50%..

2. Proses Konvergensi: Seiring bertambahnya iterasi, model menunjukkan perbaikan yang konsisten. Nilai *training loss* menyusut secara eksponensial..
3. Kinerja Akhir Pelatihan Pada akhir epoch ke-15, model berhasil menekan nilai *training loss* di angka 0.0300 dan mempertahankan *validation loss* di angka 0.0398. Akurasi validasi akhir sangat stabil di angka 98.73%. Pola penurunan *loss* yang selaras ini mengindikasikan bahwa arsitektur ConvNeXt yang dilatih bersama teknik augmentasi data mampu mengekstraksi fitur spasial secara sempurna.

Untuk melihat tren pergerakan metrik tersebut secara lebih intuitif, representasi visual dari konvergensi *loss* dan akurasi model ConvNeXt divisualisasikan pada grafik berikut.



Gambar 4. 2 Grafik Konvergensi Loss dan Akurasi Validasi pada ConvNeXt

Pola penurunan *loss* yang selaras antara data latih dan validasi pada gambar 4.2. di atas mengindikasikan bahwa arsitektur ConvNeXt mampu mengekstraksi fitur spasial bercak penyakit dengan sangat sempurna tanpa sekadar menghafal data (*memorization*).

4.2.2 Hasil Pelatihan DINOv2

Berbeda dengan ConvNeXt, arsitektur DINOv2 (*Vision Transformer*) menunjukkan dinamika pelatihan yang sedikit lebih fluktuatif pada fase evaluasi. Proses pelatihan dapat dilihat pada gambar 4.3.

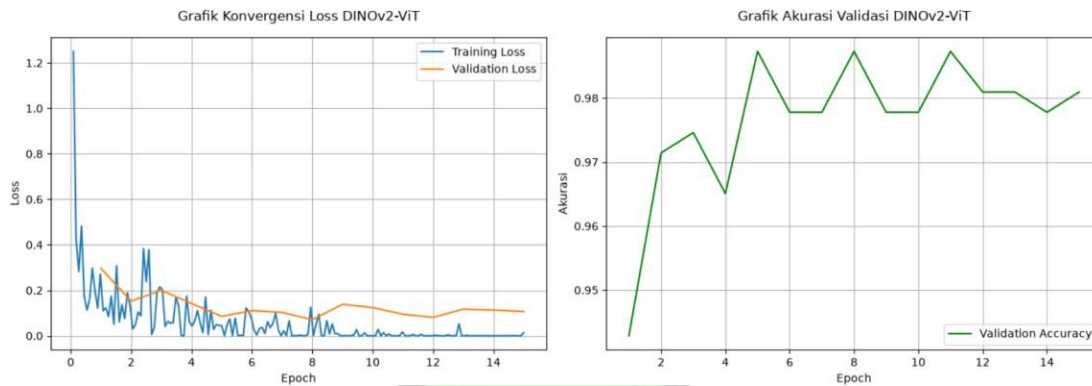
Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.271286	0.296417	0.942857	0.942620	0.948663	0.942857
2	0.132497	0.152373	0.971429	0.971454	0.972025	0.971429
3	0.215299	0.200050	0.974603	0.974595	0.975393	0.974603
4	0.061086	0.141914	0.965079	0.965121	0.965533	0.965079
5	0.044596	0.085310	0.987302	0.987302	0.987420	0.987302
6	0.078092	0.110711	0.977778	0.977790	0.978181	0.977778
7	0.000228	0.102782	0.977778	0.977717	0.977862	0.977778
8	0.126476	0.070995	0.987302	0.987302	0.987420	0.987302
9	0.000022	0.138495	0.977778	0.977776	0.978038	0.977778
10	0.000161	0.123957	0.977778	0.977776	0.977804	0.977778
11	0.016907	0.094921	0.987302	0.987299	0.987416	0.987302
12	0.000002	0.081013	0.980952	0.980995	0.981126	0.980952
13	0.000018	0.117042	0.980952	0.980995	0.981126	0.980952
14	0.000001	0.112991	0.977778	0.977822	0.977896	0.977778
15	0.013431	0.106756	0.980952	0.980995	0.981126	0.980952

Gambar 4.3 Log metrik pelatihan per-epoch pada DINOv2

Berdasarkan gambar 4.3 log perekaman performa, fenomena pembelajaran model DINOv2 dapat dirincikan sebagai berikut:

1. Pada epoch pertama, model mencatatkan *training loss* 0.2712 dan *validation loss* 0.2964, dengan tingkat akurasi validasi 94.28%.
2. Fluktuasi Tengah: Memasuki pertengahan pelatihan, model menunjukkan fluktuasi pada data validasi namun secara perlahan mampu mencapai titik konvergen.
3. Kinerja Akhir Pelatihan: Pada epoch ke-15, DINOv2 berhasil menekan *training loss* hingga angka 0.0134 dan nilai *validation loss* di angka 0.1067. Akurasi validasi akhir tercatat sebesar 98.09%. Meskipun secara umum performa ini sangat tinggi, tertahannya nilai *validation loss* saat *training loss* menyentuh nol merupakan indikasi minor dari gejala *domain overfitting* pada data latih.

Untuk memvisualisasikan tren fluktuasi dan konvergensi tersebut, representasi pergerakan *loss* dan akurasi model DINOv2 ditampilkan pada grafik berikut.



Gambar 4.4 Grafik Konvergensi Loss dan Akurasi Validasi pada DINOv2

Meskipun secara umum tingkat akurasi yang diraih sangat tinggi, pola grafik pada Gambar 4.4 yang menunjukkan tertahannya nilai *validation loss* di saat *training loss* sudah menyentuh angka nol (0.000000) merupakan indikasi minor terjadinya gejala *domain overfitting*. Model DINOv2 terindikasi terlalu "menghafal" pola data latih, sehingga tingkat kesalahannya pada data validasi tidak bisa ditekan serendah arsitektur ConvNeXt.

4.2.3 Evaluasi Metrik Kinerja

Kedua arsitektur model dievaluasi secara buta (*blind test*) menggunakan 207 citra data uji yang sama sekali belum pernah dilihat oleh model pada fase *training* dan *validation*. Citra uji ini terdiri dari 69 *Healthy Leaf*, 67 *Bacterial Leaf Disease*, dan 71 *Dried Leaf*. Kinerja model diukur secara kuantitatif menggunakan metrik *Classification Report* yang meliputi nilai *Precision*, *Recall*, dan *F1-Score*. Rincian nilai *Classification Report* untuk masing-masing model dapat dilihat pada Tabel 4.3 dan Tabel 4.4 berikut.

Tabel 4.3 *Classification Report Model ConvNeXt*

Kelas	Precision	Recall	F1-Score	Support
Healthy	0.9718	1.0000	0.9857	69
Bacterial Leaf Disease	1.0000	0.9701	0.9848	67

Dried Leaf	1.0000	1.0000	1.0000	71
Akurasi Total	0.9903 (99.03%)			207

Tabel 4.4 *Classification Report Model DINOv2*

Kelas	Precision	Recall	F1-Score	Support
Healthy	0.9014	0.9275	0.9143	69
Bacterial Leaf Disease	0.9219	0.8806	0.9008	67
Dried Leaf	0.9861	1.0000	0.9930	71
Akurasi Total	0.9372 (93.72%)			207

Berdasarkan perbandingan tabel di atas, terlihat perbedaan performa empiris yang sangat signifikan. Berkat penambahan teknik augmentasi data, arsitektur ConvNeXt mengalami lonjakan performa yang luar biasa dan secara telak mengungguli DINOv2. ConvNeXt mencatatkan akurasi pengujian nyaris sempurna sebesar 99.03%, dengan nilai *recall* sempurna (1.0000) pada kelas *Dried Leaf* dan *Healthy Leaf*, serta presisi absolut (1.0000) pada pengenalan *Bacterial Leaf Disease*. Di sisi lain, DINOv2 hanya bertahan pada akurasi 93.72%.

Metrik evaluasi kinerja tersebut dihitung secara manual menggunakan pendistribusian nilai dari *Confusion Matrix* ConvNeXt. Sebagai representasi pembuktian, berikut adalah penjabaran implementasi perhitungannya, dengan mengambil fokus pada evaluasi kelas *Bacterial Leaf Disease*.

Diketahui nilai parameter untuk kelas *Bacterial Leaf Disease* adalah:

- True Positive (TP): 65 (Citra bakteri yang ditebak benar sebagai bakteri)
- False Negative (FN): 2 (Citra bakteri yang keliru ditebak sebagai daun sehat)
- False Positive (FP): 0 (Citra daun sehat atau kering yang keliru ditebak sebagai bakteri)

- d. Total Data Uji (N): 207 citra
- e. Total Tebakan Benar Seluruh Kelas: 69 (Healthy) + 65 (Bacterial) + 71 (Dried) = 205 citra.

Maka, perhitungan metrik evaluasinya adalah sebagai berikut:

1. Perhitungan *Accuracy* Keseluruhan.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} = \frac{205}{207} = 0.9903 \approx 99\%$$

2. Perhitungan *Precision* untuk Kelas Bakteri.

$$Precision = \frac{TP}{TP+FP} = \frac{65}{65+0} = \frac{65}{65} = 1.0000 \approx 100.00\%$$

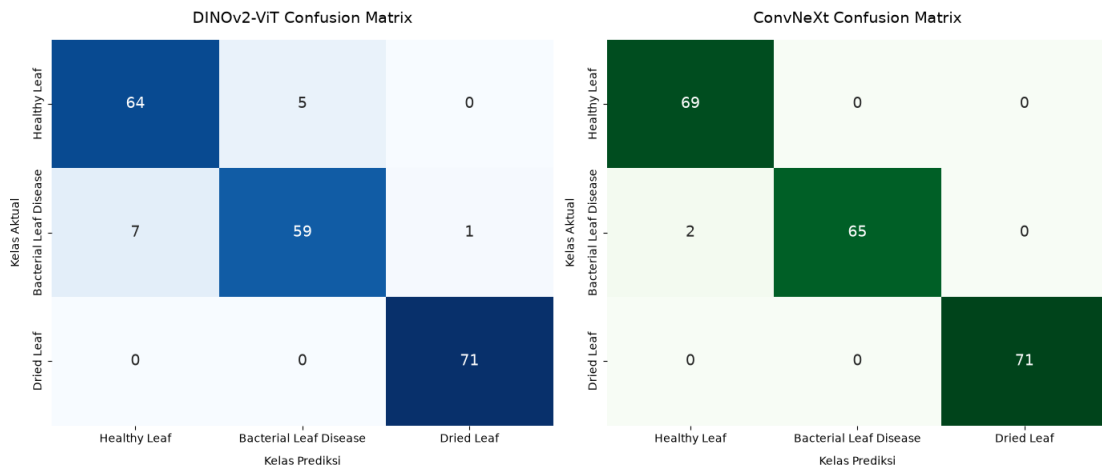
3. Perhitungan *Recall* untuk Kelas Bakteri.

$$Recall = \frac{TP}{TP+FN} = \frac{65}{65+2} = \frac{65}{67} = 0.9701 \approx 97.01\%$$

4. Perhitungan *F1-Score* untuk Kelas Bakteri.

$$F1 - score = 2x \frac{precision+recall}{precision+recall} = 2x \frac{1.0000 \times 0.9701}{1.0000 \times 0.9701} = 2x \frac{0.9701}{1.9701} = 0.9848 \approx 98.48\%$$

Untuk mengetahui secara pasti di mana letak kebingungan atau kesalahan tebakan klasifikasi pada model CNN, analisis dilanjutkan dengan melihat sebaran letak tebakan pada Matriks Konfusi (*Confusion Matrix*) berikut.



Gambar 4.5 Confusion Matrix ConvNeXt dan DINOv2

Berdasarkan *Confusion Matrix* pada Gambar 4.5, penjabaran letak prediksi kedua model adalah sebagai berikut:

1. *Confusion Matrix* ConvNeXt (CNN): Model ini menunjukkan kemampuan klasifikasi spasial yang sangat tangguh dengan tingkat akurasi total 99.03%. ConvNeXt mampu mengenali kelas *Dried Leaf* (71 citra) dan *Healthy Leaf* (69 citra) secara sempurna tanpa ada kesalahan satu pun. Pada pengenalan penyakit *Bacterial Leaf Disease*, dari 67 citra, ConvNeXt berhasil menebak 65 citra dengan benar dan hanya mengalami 2 kesalahan prediksi minor (keliru ditebak sebagai daun sehat). Hal ini membuktikan keunggulan konvolusi lokal dalam memindai dan melokalisasi bercak anomali berskala kecil secara sangat akurat.
2. *Confusion Matrix* DINOv2 (Vision Transformer): Arsitektur ini mencatatkan akurasi pengujian sebesar 93.72%. Meskipun berhasil menebak seluruh kelas *Dried Leaf* secara sempurna (71 citra), DINOv2 menunjukkan kelemahan pada kelas lain. Dari 67 citra bakteri, terdapat 8 kegagalan prediksi (7 citra keliru dianggap sebagai daun sehat dan 1 citra dianggap sebagai *Dried Leaf*). Selain itu, terdapat 5 citra *Healthy Leaf* yang keliru didiagnosis sebagai bakteri.

4.3 Pembahasan

Penelitian ini membandingkan dua pendekatan berbeda dalam mengenali penyakit daun sirih. Pendekatan pertama menggunakan arsitektur *Convolutional Neural Network* (CNN) pada model ConvNeXt yang fokus mencari pola pada area kecil (lokal). Pendekatan kedua menggunakan *Vision Transformer* (DINOv2) yang memproses gambar secara keseluruhan (global). Dari hasil pengujian, model ConvNeXt terbukti lebih unggul dibandingkan DINOv2.

Keberhasilan ConvNeXt berasal dari cara kerjanya yang menggunakan filter konvolusi. Filter ini bertugas memindai gambar secara detail pada area-area tertentu saja. Cara ini membuat model mampu menemukan titik bercak infeksi bakteri yang ukurannya sangat kecil dengan tepat, tanpa terganggu oleh luasnya bagian daun lain yang warnanya masih hijau dan sehat.

Sebaliknya, cara kerja DINOv2 yang melihat gambar secara keseluruhan justru menjadi kelemahan tersendiri. Model ini sering kali kesulitan saat harus mendeteksi penyakit yang ukurannya kecil. Pola bercak penyakit tersebut sering kali terabaikan atau "tenggelam" karena sistem lebih banyak melihat area daun sehat yang ukurannya jauh lebih besar dan mendominasi gambar.

Perbedaan kinerja kedua model ini terlihat sangat jelas dari hasil pengujian. ConvNeXt meraih akurasi sebesar 99.03%, jauh melampaui DINOv2 yang hanya mendapat 93.72%. Keakuratan ini juga dibuktikan oleh gambar visual *heatmap* Grad-CAM, di mana fokus tebakan ConvNeXt terlihat pas mengenai titik penyakit, sedangkan fokus DINOv2 tampak tersebar. Hasil ini membuktikan bahwa model ConvNeXt sangat tangguh dan siap dikembangkan menjadi sistem deteksi otomatis untuk membantu sektor pertanian.

Untuk memvalidasi signifikansi hasil, performa model dikomparasikan dengan studi terdahulu. Kinerja tangguh ConvNeXt pada penelitian ini (akurasi 99.03% dari 2.307 citra daun sirih) menunjukkan konsistensi pengenalan fitur lokal yang sejalan dengan penelitian Murugesan dkk. (2025), yang juga sukses mencapai akurasi 98.92%

menggunakan ConvNeXt pada dataset masif 12.865 citra penyakit daun tanaman. Di sisi lain, capaian DINOv2 pada penelitian ini (akurasi 93.72%) selaras dengan temuan Hussien dkk. (2025) yang meraih akurasi 95.00% menggunakan DINOv2 untuk klasifikasi 4.000 citra penyakit retina.

Melalui komparasi ini, model ConvNeXt terbukti tidak hanya memiliki kecepatan konvergensi yang sangat prima dan menekan kurva kerugian (*loss*) mendekati nol hanya dalam 15 iterasi, tetapi juga membuktikan bahwa ketajaman konvolusi lokalnya jauh lebih superior dan efisien dibandingkan mekanisme global ViT dalam memetakan anomali tekstur tipis, bahkan pada skala dataset yang lebih terbatas.



BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan penelitian serta analisis yang telah dilakukan, maka dapat diangkat kesimpulan sebagai berikut:

1. Penerapan model *deep learning* Convolutional Neural Network (ConvNeXt) dan Vision Transformer (DINOv2) terbukti memiliki paradigma yang sangat berbeda dalam mengenali pola penyakit daun sirih. Arsitektur ConvNeXt menerapkan operasi konvolusi yang berfokus pada ekstraksi fitur lokal, sehingga sangat andal dalam memindai detail tekstur mikroskopis dari luka penyakit. Sebaliknya, arsitektur DINOv2 menerapkan mekanisme *global self-attention* yang memproses konteks citra secara menyeluruh, di mana pendekatan ini justru menjadi kelemahan tersendiri karena model rentan mengabaikan anomali berskala kecil yang tertutup oleh warna dominan daun sehat.
2. Performa model Convolutional Neural Network (ConvNeXt) terbukti jauh lebih unggul, seimbang, dan akurat dibandingkan Vision Transformer (DINOv2). ConvNeXt berhasil meraih akurasi pengujian sebesar 99.03%, yang didukung oleh capaian luar biasa pada metrik lainnya, yaitu nilai *recall* sempurna (1.0000) pada kelas *Dried Leaf* dan *Healthy Leaf*, presisi absolut (1.0000) pada kelas daun bakteri, serta keseimbangan F1-Score yang sangat memukau di seluruh kelas yang berkisar antara 0.98 hingga 1.00. Di sisi lain, model DINOv2 hanya mencatatkan akurasi sebesar 93.72% dengan kelemahan yang cukup jelas pada pengenalan daun bakteri. Hal ini dibuktikan dengan rendahnya metrik *recall* DINOv2 pada kelas bakteri (0.8806) yang turut menarik turun nilai F1-Score-nya menjadi (0.9008), sehingga menyimpulkan bahwa arsitektur CNN memberikan hasil klasifikasi multikelas yang jauh lebih presisi dan dapat diandalkan.

3. Hasil evaluasi visual menggunakan peta *heatmap* pada Grad-CAM membuktikan secara empiris letak keunggulan dan kelemahan dari masing-masing model. ConvNeXt mampu memusatkan atensinya secara tajam dan konsisten tepat pada titik pusat lesi penyakit (bercak bakteri). Sebaliknya, pola atensi DINOv2 terbukti terpecah dan tersebar tidak beraturan ke seluruh permukaan daun sehat, yang mengonfirmasi bahwa arsitektur berbasis *Transformer* ini memiliki "titik buta" dalam melokalisasi target spesifik pada citra daun.
4. Secara keseluruhan, arsitektur CNN modern (ConvNeXt) terbukti jauh lebih superior, stabil, dan dapat diinterpretasikan secara visual (*explainable*) dibandingkan model *Self-Supervised* Vision Transformer (DINOv2) untuk kasus pengenalan anomali daun sirih. Oleh karena itu, pendekatan konvolusi lokal tetap menjadi arsitektur yang paling direkomendasikan untuk diimplementasikan sebagai dasar pengembangan alat bantu deteksi otomatis di sektor pertanian

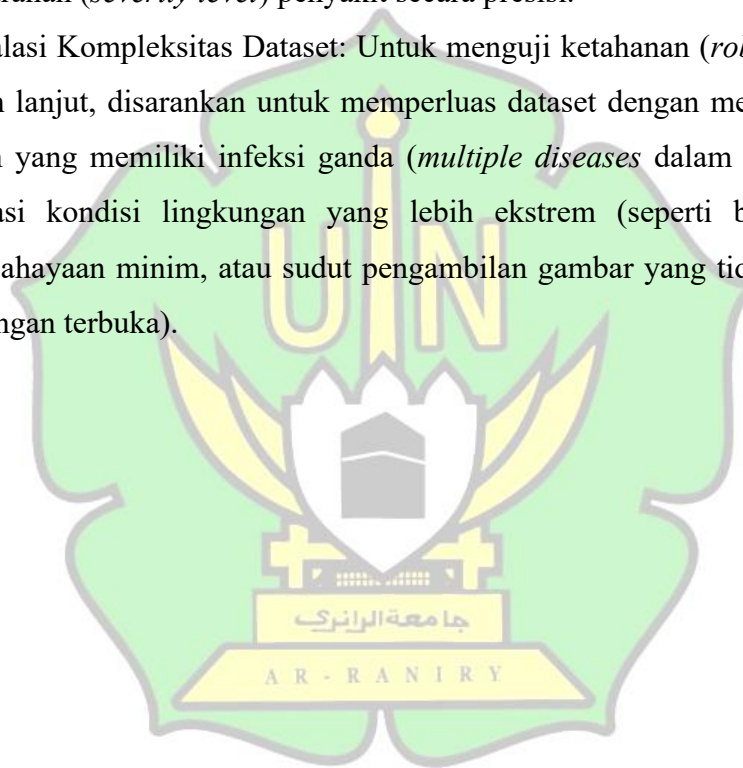
5.2 Saran

Penelitian ini kemungkinan masih memiliki kesempatan untuk dilanjutkan atau dikembangkan ke depannya, berikut saran saran yang dapat diajukan sebagai berikut:

1. Pengembangan Arsitektur Hibrida (CNN-ViT): Mengingat arsitektur ConvNeXt (CNN) lebih unggul dibandingkan dengan arsitektur DINOv2 (ViT), penelitian selanjutnya sangat disarankan untuk membangun arsitektur hybrid atau penggabungan model. Kombinasi ini memungkinkan akan membantu mengatasi titik buta" pada model murni *Transformer* sekaligus mempertahankan ketajaman deteksi anomali mikroskopis.
2. Implementasi Sistem Berbasis Mobile Terintegrasi: Untuk meningkatkan nilai guna penelitian di sektor praktis, model klasifikasi ConvNeXt yang telah terbukti memiliki akurasi tinggi dapat di *deploy* ke dalam aplikasi berbasis *mobile* atau web. Hal ini memungkinkan sistem untuk digunakan secara *real-time* menggunakan kamera *smartphone* oleh para petani sirih, khususnya dalam

mendukung efisiensi pemantauan hasil komoditas pertanian lokal maupun global.

3. Pengembangan Menuju Estimasi Tingkat Keparahan Penyakit: Penelitian ini saat ini berfokus pada ranah klasifikasi citra (*image classification*). Penelitian selanjutnya dapat mengeksplorasi algoritma deteksi objek atau segmentasi instans (*instance segmentation*) seperti YOLO atau Mask R-CNN. Tujuannya bukan hanya untuk mengenali jenis penyakit, tetapi juga untuk melokalisasi dan menghitung persentase luas area daun yang terinfeksi guna menentukan tingkat keparahan (*severity level*) penyakit secara presisi.
4. Eskalasi Kompleksitas Dataset: Untuk menguji ketahanan (*robustness*) model lebih lanjut, disarankan untuk memperluas dataset dengan mengakuisisi citra daun yang memiliki infeksi ganda (*multiple diseases* dalam satu daun) atau variasi kondisi lingkungan yang lebih ekstrem (seperti bayangan tebal, pencahayaan minim, atau sudut pengambilan gambar yang tidak beraturan di lapangan terbuka).



DAFTAR PUSTAKA

- Ahmed, A., Ahmed, J., Chakrabarty, T., Muqit, A., & Asad-ud-doullah, M. (2021). An investigation into disease problems of khasia paan (Piper betle) in the sylhet region. *Bangladesh Journal Plant Pathology*, 37(1&2), 87–92. https://www.researchgate.net/profile/Junaed-Ahmed-2/publication/360540894%5C_An%5C_investigation%5C_into%5C_disease%5C_problems%5C_of%5C_Khasia%5C_Paan%5C_Piper%5C_betle%5C_i_n%5C_the%5C_Sylhet%5C_region/links/627cd38b107cae29199ee94b/An-investigation-into
- Alnuaimi, A. F. A. H., & Albaldawi, T. H. K. (2024). An overview of machine learning classification techniques. *BIO Web of Conferences*, 97, 1–24. <https://doi.org/10.1051/bioconf/20249700133>
- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. <https://doi.org/10.1186/s40537-021-00444-8>
- Bhowmik, A. C., Ahad, M. T., Emon, Y. R., Ahmed, F., Song, B., & Li, Y. (2024). A customised vision transformer for accurate detection and classification of Java Plum leaf disease. *Smart Agricultural Technology*, 8(July), 100500. <https://doi.org/10.1016/j.atech.2024.100500>
- Böhle, M., Fritz, M., & Schiele, B. (2021). Convolutional Dynamic Alignment Networks for Interpretable Classifications. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 10024–10033. <https://doi.org/10.1109/CVPR46437.2021.00990>
- Brondolo, F., & Beaussant, S. (2025). DINOv2 rocks geological image analysis: Classification, segmentation, and interpretability. *Journal of Rock Mechanics*

and *Geotechnical Engineering*, 17(11), 6853–6867.
<https://doi.org/10.1016/j.jrmge.2025.01.057>

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale. *ICLR 2021 - 9th International Conference on Learning Representations*.

Erda, G., Sari, S. F., Astuti, S. D., Thearas, L., & Yolanda, A. M. (2025). Perbandingan Kinerja Machine Learning Dalam Klasifikasi Minat Mbkm Mahasiswa Unri. *Jurnal Informatika Dan Rekayasa ...*, 8(2), 188–198.

Hassan, A., Refaat, M., & Hemeida, A. (2022). Image classification based deep learning: A Review. *Aswan University Journal of Sciences and Technology*, 2(1), 11–35. <https://doi.org/10.21608/aujst.2022.259887>

Hridoy, R. H., Tarek Habib, M., Sadekur Rahman, M., & Uddin, M. S. (2022). Deep Neural Networks-Based Recognition of Betel Plant Diseases by Leaf Image Classification. *Lecture Notes on Data Engineering and Communications Technologies*, 116, 227–241. https://doi.org/10.1007/978-981-16-9605-3_16

Hu, L., Chen, J., Vaughan, J., Yang, H., Wang, K., Sudjianto, A., & Nair, V. N. (2020). *Supervised machine learning techniques: An overview with applications to banking corporate model risk [Technical report]. Wells Fargo*. 1–29.

Hussien, A., Elkhateb, A., Saeed, M., Elsabawy, N. M., Elnakeeb, A. E., & Elrashidy, N. (2025). Explainable self-supervised learning for medical image diagnosis based on DINO V2 model and semantic search. *Scientific Reports*, 15(1), 1–22. <https://doi.org/10.1038/s41598-025-15604-6>

Jackulin, C., & Murugavalli, S. (2022). A comprehensive review on detection of plant disease using machine learning and deep learning approaches. *Measurement: Sensors*, 24, 1–37. <https://doi.org/10.1016/j.measen.2022.100441>

- Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2022-June*, 11966–11976. <https://doi.org/10.1109/CVPR52688.2022.01167>
- Murugesan, S., Chinnadurai, J., Srinivasan, S., Mathivanan, S. K., Chandan, R. R., & Moorthy, U. (2025). Robust multiclass classification of crop leaf diseases using hybrid deep learning and Grad-CAM interpretability. *Scientific Reports, 15*(1), 1–22. <https://doi.org/10.1038/s41598-025-14847-7>
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P. Y., Li, S. W., Misra, I., Rabbat, M., Sharma, V., ... Bojanowski, P. (2024). DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research, 2024*, 1–32.
- Pakiding, D., Selao, A., & Wahyuddin, W. (2025). Implementasi Computer Vision dalam Mendeteksi Penyakit pada Tanaman Cabai dan Tomat Menggunakan Algoritma Convolutional Neural Networks. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 5*(3), 841–850. <https://doi.org/10.57152/malcom.v5i3.1989>
- Shehan, A., Anjali, S., Sanjula, L., Lavanya, M., Thelijjagoda, S., & Krishara, J. (2023). Neuroassist ADHD Analyzer: A Smart Application for Recognizing Attention Deficit Hyperactivity Disorder (ADHD) Levels in Children. *Proceedings of the 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems, ICSES 2023*, 710–721. <https://doi.org/10.1109/ICSES60034.2023.10465399>
- Urrea, C., & Vélez, M. (2025). Advances in Deep Learning for Semantic Segmentation of Low-Contrast Images: A Systematic Review of Methods, Challenges, and Future Directions. *Sensors, 25*(7). <https://doi.org/10.3390/s25072043>

- Utami, D., Meale, S. J., & Young, A. J. (2022). A Pan-Global Study of Bacterial Leaf Spot of Chilli Caused by *Xanthomonas* spp. *Plants*, *11*(17), 1–19. <https://doi.org/10.3390/plants11172291>
- Wahid, J. A., Mingliang, X., Ayoub, M., Hussain, S., Li, L., & Shi, L. (2024). A hybrid ResNet-ViT approach to bridge the global and local features for myocardial infarction detection. *Scientific Reports*, *14*(1), 1–17. <https://doi.org/10.1038/s41598-024-54846-8>
- Wang, X., Wang, Y., Zhao, J., & Niu, J. (2023). ECA-ConvNeXt: A Rice Leaf Disease Identification Model Based on ConvNeXt. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2023-June*, 6235–6243. <https://doi.org/10.1109/CVPRW59228.2023.00663>
- Wang, Y., Deng, Y., Zheng, Y., Chattopadhyay, P., & Wang, L. (2025). Vision Transformers for Image Classification: A Comparative Survey. *Technologies*, *13*(1), 1–32. <https://doi.org/10.3390/technologies13010032>
- Wong, G. (2025). A comparative study of Convolutional Neural Network and Vision Transformer models as classifiers of east Asian traditional clothing. *Journal of High School Science*, *9*(3), 345–369. <https://doi.org/10.64336/001c.143597>
- Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. In *Artificial Intelligence Review* (Vol. 57, Issue 4). Springer Netherlands. <https://doi.org/10.1007/s10462-024-10721-6>