

Analisis Karakteristik Visual Dataset dan Pola Kesalahan Deteksi untuk Evaluasi Generalisasi YOLOv8 pada Penyakit Daun Cabai

Fajar Wahyu Hanafi^{1*}, Rudi Susanto², Herliyani Hasanah³, Wiji Lestari⁴, Triyono⁵

¹Teknik Informatika
Universitas Duta Bangsa Surakarta
^{1*}220103057@mhs.udb.ac.id

²Teknik Komputer
Universitas Duta Bangsa Surakarta
²rudi_susanto@udb.ac.id

³Teknik Informatika
Universitas Duta Bangsa Surakarta
³herliyani_hasanah@udb.ac.id

⁴Sistem Informasi
Universitas Duta Bangsa Surakarta
⁴wiji_lestari@udb.ac.id

⁵Teknik Informatika
Universitas Duta Bangsa Surakarta
⁵triyono@udb.ac.id

Abstrak— Model deteksi penyakit daun berbasis deep learning umumnya mencapai akurasi tinggi pada dataset dengan karakteristik serupa, namun performanya sering kali menurun pada citra lapangan akibat perbedaan karakteristik visual antara data pelatihan dan pengujian (domain shift). Penelitian ini bertujuan menganalisis perbedaan karakteristik visual dataset publik dan lapangan serta bagaimana perbedaan tersebut tercermin pada performa dan pola kesalahan deteksi YOLOv8, menggunakan dataset penyakit daun cabai (3.152 citra publik dan 800 citra lapangan). Metode karakteristik dataset dianalisis menggunakan statistik deskriptif terhadap brightness, sharpness, ukuran bounding box, dan jumlah objek per citra, dilanjutkan dengan evaluasi cross-domain dua model YOLOv8 (Model A dan Model B) pada data uji publik dan lapangan, serta analisis pola kesalahan berbasis Intersection over Union (IoU) terhadap false negative, false positive, dan misklasifikasi. Hasil penelitian menunjukkan dataset lapangan memiliki kecerahan dan ketajaman lebih rendah serta ukuran objek lebih kecil dibandingkan dataset publik, diikuti penurunan performa model pada pengujian lintas domain, terutama pada kelas *keriting daun* dan *virus kuning*. Analisis pola kesalahan memperlihatkan false negative lebih sering terjadi pada objek kecil dan kondisi crowding, sedangkan false positive serta misklasifikasi berkaitan dengan kemiripan karakteristik visual antar kelas. Temuan ini menunjukkan bahwa analisis karakteristik dataset dan pola kesalahan memberikan pemahaman lebih menyeluruh mengenai domain shift serta menjadi dasar pengembangan model dengan generalisasi lebih baik pada kondisi lapangan.

Kata kunci— YOLOv8, domain shift, deteksi penyakit daun cabai, generalisasi model, analisis kesalahan.

Abstract— Deep learning-based leaf disease detection models typically achieve high accuracy on similar-characteristic datasets, but performance declines on field images due to visual differences between training and testing data (domain shift). This study aims to analyze visual characteristic differences between public and field datasets and how these are reflected in the performance and error patterns of YOLOv8 detection, using a chili leaf disease dataset (3,152 public and 800 field images). Method dataset characteristics were analyzed using descriptive statistics of brightness, sharpness, bounding box size, and object count per image, followed by cross-domain evaluation of two YOLOv8 models (Model A and Model B) on public and field test data, and error pattern analysis based on Intersection over Union (IoU) for false negatives, false positives, and misclassifications. Results show the field dataset has lower brightness and sharpness and smaller objects than the public dataset, followed by a performance decline under cross-domain testing, particularly for the leaf-curl and yellow-leaf-virus classes. Error analysis shows false negatives occur more often in small objects and crowded conditions, while false positives and misclassifications relate to visual similarity between disease classes. These findings provide a more thorough understanding of domain shift and serve as a basis for developing models with better generalization under field conditions.

Keywords— YOLOv8, domain shift, chili leaf disease detection, model generalization, error analysis.

I. PENDAHULUAN

Cabai rawit (*Capsicum frutescens* L.) merupakan komoditas hortikultura strategis di Indonesia yang berperan penting dalam stabilitas harga pangan nasional dan kehidupan petani [1]. Budidaya cabai secara konsisten terancam

oleh berbagai penyakit daun, termasuk virus kuning (*virus kuning*), keriting daun (*keriting daun*), dan bercak daun (*bercak daun*), yang merusak kapasitas fotosintesis dan menurunkan kualitas hasil panen [2]. Oleh karena itu, identifikasi penyakit secara akurat dan tepat waktu menjadi langkah penting dalam

mendukung pengendalian penyakit secara efektif [3].

Seiring berkembangnya teknologi computer vision, pendekatan berbasis *deep learning*, khususnya penggunaan *You Only Look Once* (YOLO), telah banyak dimanfaatkan untuk mendukung deteksi otomatis penyakit tanaman dari citra daun karena mampu memberikan keseimbangan antara kecepatan inferensi dan akurasi deteksi [4][5]. Meskipun demikian, performa yang tinggi tersebut umumnya diperoleh pada data yang memiliki karakteristik serupa dengan data pelatihan [6]. Ketika diterapkan pada citra lapangan dengan kondisi visual yang lebih beragam, performa model cenderung menurun sehingga kemampuan generalisasinya menjadi tantangan dalam implementasi di dunia nyata [7].

Fenomena tersebut dikenal sebagai domain shift, atau kondisi ketika karakteristik visual antara data pelatihan dan data yang digunakan pada tahap pengujian berbeda sehingga kemampuan model untuk mengenali objek pada domain baru menjadi menurun [8]. Dalam konteks deteksi penyakit tanaman, perbedaan tersebut dapat berupa variasi intensitas pencahayaan, kompleksitas latar belakang, sudut dan jarak pengambilan gambar, tingkat ketajaman citra, maupun ukuran objek penyakit antara dataset publik dan citra yang diperoleh langsung di lapangan [9].

Penelitian sebelumnya oleh Zhu dkk. telah melakukan kajian terhadap permasalahan domain shift pada detektor berbasis YOLO, dengan menunjukkan bahwa performa detektor yang dilatih pada satu domain cenderung menurun secara signifikan ketika diterapkan pada domain lain, sehingga diperlukan strategi adaptasi domain untuk mengurangi kesenjangan distribusi antar domain tersebut. Penelitian tersebut berfokus pada pengembangan metode adaptasi berbasis fitur pada arsitektur YOLO untuk meningkatkan performa deteksi lintas domain secara umum. Namun, penelitian tersebut belum mengkaji secara spesifik karakteristik visual dataset maupun pola kesalahan deteksi yang menyertai fenomena domain shift, khususnya dalam konteks deteksi penyakit daun tanaman

[10]. Sementara itu, Yang dkk. mengembangkan kerangka kerja cross-domain few-shot learning untuk identifikasi penyakit tanaman pada kondisi lapangan dengan memanfaatkan pembelajaran lintas domain pada skenario data terbatas. Penelitian tersebut menunjukkan bahwa peningkatan kemiripan distribusi antar-domain berkontribusi terhadap peningkatan akurasi klasifikasi, namun evaluasi yang dilakukan masih berfokus pada perbandingan performa model antar-domain dan belum melakukan analisis kuantitatif terhadap karakteristik visual dataset maupun hubungannya dengan pola kesalahan prediksi model [9].

Berdasarkan kondisi tersebut, penelitian ini bertujuan untuk menganalisis perbedaan karakteristik visual antara dataset publik dan dataset lapangan serta mengevaluasi bagaimana perbedaan tersebut tercermin pada kemampuan generalisasi dan pola kesalahan deteksi model YOLO. Analisis dilakukan melalui tiga tahapan utama, yaitu karakterisasi dataset menggunakan statistik deskriptif, evaluasi lintas domain terhadap model yang telah dilatih sebelumnya, serta analisis pola kesalahan deteksi berdasarkan hasil inferensi model. Melalui pendekatan tersebut, penelitian ini diharapkan dapat memberikan pemahaman yang lebih komprehensif mengenai hubungan antara karakteristik dataset, kemampuan generalisasi model, dan pola kesalahan deteksi, sehingga dapat menjadi dasar dalam pengembangan dataset maupun strategi pelatihan yang lebih representatif untuk implementasi pada kondisi lapangan.

II. METODOLOGI PENELITIAN

A. Dataset dan Model

Penelitian menggunakan dua dataset yang diperoleh melalui platform Roboflow, yaitu dataset publik 3.512 citra dan dataset lapangan hasil pengumpulan langsung 800 citra, yang keduanya mencakup empat kelas: *bercak_daun*, *daun_sehat*, *keriting_daun*, dan *virus_kuning*. Kedua dataset telah melalui proses pra-pemrosesan yang sama, yaitu penyeragaman resolusi citra menjadi 640×640 piksel.

Penelitian ini menggunakan dua model YOLOv8 yang diperoleh dari penelitian sebelumnya, yaitu Model A (dilatih menggunakan dataset publik) dan Model B (dilatih menggunakan dataset publik dengan augmentasi). Kedua model dipilih karena penelitian ini bertujuan mengevaluasi kemampuan generalisasi model yang dilatih pada domain publik ketika diterapkan pada domain yang berbeda. Model hasil sequential fine-tuning tidak digunakan karena bobot model telah diadaptasi menggunakan dataset lapangan sehingga tidak lagi merepresentasikan generalisasi dari domain public.

B. Analisis Karakteristik Dataset

Karakteristik visual dan karakteristik objek pada kedua dataset dianalisis melalui empat variabel, yaitu brightness (rata-rata intensitas piksel citra dalam ruang warna grayscale), sharpness (dihitung menggunakan variance of Laplacian), luas bounding box relatif (proporsi luas objek terhadap luas keseluruhan citra), dan jumlah objek per citra. Untuk setiap variabel dihitung statistik deskriptif berupa mean, median, standar deviasi, nilai minimum, dan nilai maksimum, kemudian divisualisasikan dalam bentuk histogram dan boxplot untuk membandingkan distribusi antara dataset publik dan dataset lapangan, baik secara keseluruhan maupun per kelas. Metrik evaluasi yang dikumpulkan meliputi Precision, Recall, F1-score, mAP@0.5, mAP@0.5:0.95, serta Average Precision (AP) untuk setiap kelas penyakit.

C. Evaluasi Generalisasi Model

Model A dan Model B dievaluasi secara cross-domain menggunakan skema inferensi (bukan pelatihan ulang) pada test split dataset publik dan test split dataset lapangan, menggunakan framework Ultralytics YOLOv8. Metrik evaluasi yang dikumpulkan meliputi Precision, Recall, F1-score, mAP@0.5, mAP@0.5:0.95, serta Average Precision (AP) untuk setiap kelas penyakit.

Precision menyatakan proporsi prediksi positif yang benar, dirumuskan sebagai:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Recall menyatakan proporsi objek yang berhasil terdeteksi, dirumuskan sebagai:

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

F1-score menyatakan rata-rata harmonik Precision dan Recall, dirumuskan sebagai:

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

AP menyatakan luas area di bawah kurva Precision-Recall, dirumuskan sebagai:

$$AP = \int_0^1 P(R) dR \quad (4)$$

mAP@0.5 menyatakan rata-rata AP seluruh kelas pada IoU 0,5, dirumuskan sebagai:

$$mAP@0.5 = \frac{1}{N} \sum_{i=1}^N AP_i \quad (5)$$

mAP@0.5:0.95 menyatakan rata-rata mAP pada ambang batas IoU 0,5 hingga 0,95 dengan interval 0,05, dirumuskan sebagai:

$$mAP@0.5:0.95 = \frac{1}{10} \sum_{i=0.5, \dots, 0.95} mAP@t \quad (6)$$

dengan TP, FP, dan FN berturut-turut menyatakan jumlah deteksi benar, deteksi salah, dan objek yang gagal terdeteksi.

D. Analisa Pola Kesalahan

Tahap selanjutnya adalah menganalisis pola kesalahan deteksi berdasarkan hasil inferensi pada dataset lapangan. Prediksi model dicocokkan dengan ground truth menggunakan Intersection over Union (IoU) dengan ambang batas 0,5 [11]. Prediksi yang memiliki kelas dan lokasi sesuai dengan ground truth dikategorikan sebagai deteksi benar, prediksi dengan lokasi sesuai namun kelas berbeda dikategorikan sebagai misclassification, prediksi tanpa pasangan ground truth dikategorikan sebagai false positive, sedangkan objek ground truth yang tidak memiliki pasangan prediksi dikategorikan sebagai false negative. Selanjutnya, setiap kategori kesalahan diverifikasi melalui inspeksi visual terhadap contoh citra untuk mengidentifikasi pola kesalahan yang sering muncul pada masing-masing domain.

E. Sintesis

Tahap terakhir menyusun sintesis yang menghubungkan hasil karakterisasi dataset, evaluasi lintas domain, dan analisis pola kesalahan ke dalam satu kerangka interpretasi. Sintesis disajikan dalam bentuk tabel yang

memuat karakteristik dataset, hasil evaluasi, pola kesalahan, dan interpretasi observasional.

III. HASIL DAN PEMBAHASAN

A. Karakteristik Visual Dataset

Pada tahap ini, dilakukan perbandingan karakteristik brightness dan sharpness antara dataset publik dan dataset lapangan untuk mengidentifikasi apakah terdapat perbedaan distribusi visual antara kedua domain. Hasil Statistik deskriptif brightness disajikan pada Tabel I, sedangkan statistik deskriptif sharpness disajikan pada Tabel II.

Tabel I. Statistik Deskriptif Brightness

Dataset	Mean	Median	Std	Min	Max
Publik	126,84	126,21	14,17	63,57	211,20
Lapangan	119,88	119,26	10,41	73,04	159,04

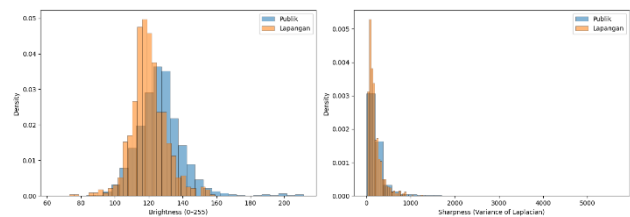
Berdasarkan Tabel I, dataset publik memiliki rata-rata brightness yang lebih tinggi dibandingkan dataset lapangan, yaitu 126,84 dan 119,88, dengan sebaran intensitas yang juga lebih lebar (standar deviasi 14,17 dibandingkan 10,41). Temuan ini menunjukkan bahwa kedua dataset dikumpulkan pada kondisi pencahayaan yang berbeda sehingga distribusi intensitas piksel yang diterima model tidak sepenuhnya seragam. Kondisi tersebut konsisten dengan penelitian referensi [8] yang menjelaskan bahwa variasi pencahayaan merupakan salah satu karakteristik visual yang membedakan citra hasil kurasi dan citra lapangan.

Tabel II. Statistik Deskriptif Sharpness

Dataset	Mean	Median	Std	Min	Max
Publik	235,59	169,78	336,47	4,92	5672,31
Lapangan	187,18	138,79	159,59	17,95	1272,91

Tabel II menunjukkan bahwa dataset publik memiliki nilai sharpness yang lebih tinggi dengan median masing-masing sebesar 169,78 dibandingkan dataset lapangan 138,79. Karena distribusi sharpness pada kedua dataset bersifat *right-skewed*, median digunakan sebagai ukuran utama, sedangkan nilai rata-rata (235,59 berbanding 187,18) hanya digunakan sebagai pembanding. Perbedaan koefisien variasi sharpness antara dataset publik (1,43) dan dataset lapangan (0,85) juga cukup besar, meskipun kedua dataset memiliki resolusi citra yang sama (640×640 piksel), sehingga variasi ini bukan

disebabkan oleh perbedaan resolusi, melainkan oleh perbedaan gaya pengambilan gambar antar dataset [8].



Gambar 1. Histogram distribusi brightness dan sharpness pada dataset publik dan dataset lapangan.

Histogram pada Gambar 1 menunjukkan bahwa distribusi sharpness pada kedua dataset sama-sama bersifat *right-skewed*, namun dataset publik memiliki penyebaran nilai yang lebih luas dibandingkan dataset lapangan. Pengamatan terhadap beberapa citra dengan nilai sharpness ekstrem menunjukkan bahwa citra lapangan dengan nilai sharpness rendah umumnya memiliki latar belakang yang lebih buram, sedangkan citra publik dengan nilai sharpness tinggi banyak dijumpai pada latar belakang yang tetap berada dalam fokus. Temuan ini mengindikasikan bahwa variasi sharpness yang diamati lebih berkaitan dengan perbedaan kondisi akuisisi citra dan komposisi latar belakang dibandingkan tingkat ketajaman objek daun yang dideteksi. Interpretasi tersebut konsisten dengan penelitian sebelumnya yang melaporkan bahwa variasi kondisi pengambilan citra turut memengaruhi karakteristik visual dataset pada deteksi penyakit tanaman [7].

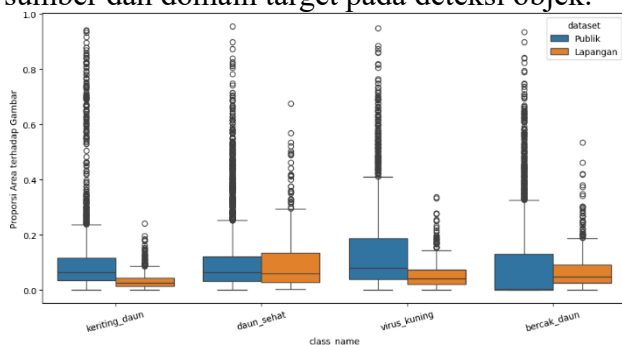
B. Karakteristik Objek

Pada tahap ini dilakukan perbandingan karakteristik objek dengan membandingkan luas bounding box dan jumlah objek per gambar. Tabel III menyajikan statistik deskriptif luas bounding box relatif per kelas, sedangkan Tabel IV menyajikan statistik jumlah objek per gambar.

Tabel III. Statistik Deskriptif Luas Bounding Box Relatif Per Kelas

Dataset	Kelas	Mean	Median	Std	Min	Max
Lapangan	<i>bercak daun</i>	0,0702	0,0488	0,0674	0,0020	0,5349
Lapangan	<i>daun sehat</i>	0,1023	0,0590	0,1129	0,0043	0,6758
Lapangan	<i>keriting daun</i>	0,0348	0,0259	0,0307	0,0018	0,2411
Lapangan	<i>virus kuning</i>	0,0564	0,0420	0,0510	0,0000	0,3376
Publik	<i>bercak daun</i>	0,0921	0,0022	0,1600	0,0000	0,9354
Publik	<i>daun sehat</i>	0,1064	0,0640	0,1242	0,0000	0,9547
Publik	<i>keriting daun</i>	0,1026	0,0640	0,1308	0,0001	0,9425
Publik	<i>virus kuning</i>	0,1430	0,0800	0,1543	0,0000	0,9484

Berdasarkan Tabel III, median luas bounding box kelas *keriting_daun* pada dataset lapangan (0,0259) lebih kecil dibandingkan dataset publik (0,0640), sedangkan pada kelas *virus_kuning* median luas bounding box juga lebih rendah (0,0420 dibandingkan 0,0800). Sebaliknya, kelas *daun_sehat* menunjukkan ukuran objek yang relatif serupa pada kedua dataset. Pada kelas *bercak_daun*, dataset publik memiliki median yang sangat kecil (0,0022) tetapi rata-rata yang jauh lebih tinggi (0,0921), mengindikasikan adanya variasi ukuran objek yang sangat besar dalam kelas tersebut. Secara keseluruhan, hasil ini menunjukkan bahwa dataset publik memiliki keragaman ukuran objek yang lebih tinggi dibandingkan dataset lapangan. Temuan ini konsisten dengan referensi [8] yang melaporkan bahwa variasi skala objek merupakan salah satu karakteristik visual yang membedakan domain sumber dan domain target pada deteksi objek.

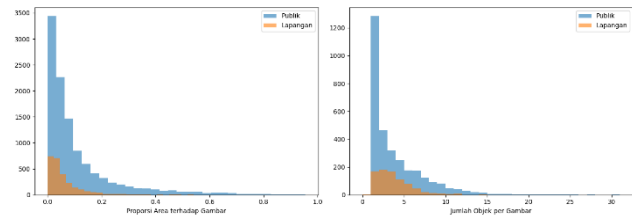


Gambar 2. Perbandingan ukuran bounding box per kelas antara dataset publik dan dataset lapangan.

Perbedaan ukuran objek antar kelas juga terlihat pada Gambar 2. Sebagian besar kelas pada dataset lapangan memiliki bounding box yang terkonsentrasi pada ukuran kecil hingga sedang, sedangkan dataset publik menunjukkan penyebaran yang lebih luas dengan keberadaan objek berukuran jauh lebih besar pada beberapa kelas. Pola ini memperkuat bahwa variasi ukuran objek pada dataset publik lebih tinggi dibandingkan dataset lapangan.

Tabel IV. Statistik Deskriptif Jumlah Objek Per Gambar

Dataset	Mean	Median	Std	Min	Max
Lapangan	3,20	3,0	2,06	1	14
Publik	3,45	2,0	3,37	1	30



Gambar 3. Histogram distribusi luas bounding box dan jumlah objek per gambar pada dataset publik dan dataset lapangan.

Tabel IV menunjukkan bahwa rata-rata jumlah objek per gambar pada kedua dataset relatif serupa, yaitu 3,20 pada dataset lapangan dan 3,45 pada dataset publik. Meskipun demikian, dataset publik memiliki variasi jumlah objek yang lebih besar, tercermin dari standar deviasi yang lebih tinggi (3,37 dibandingkan 2,06) serta jumlah objek maksimum yang mencapai 30, sedangkan dataset lapangan hanya mencapai 14 objek. Histogram pada Gambar 3 memperlihatkan bahwa sebagian besar citra lapangan terkonsentrasi pada rentang 1–5 objek per gambar, sedangkan dataset publik memiliki distribusi yang lebih lebar hingga lebih dari 20 objek pada sebagian kecil citra. Kondisi ini mengindikasikan bahwa komposisi objek pada dataset publik lebih beragam, sementara dataset lapangan cenderung memiliki jumlah objek yang lebih konsisten antar citra.

C. Evaluasi Generalisasi Model

Pada subbab ini dilakukan evaluasi kemampuan generalisasi Model A dan Model B melalui pengujian lintas domain menggunakan test split dataset publik dan dataset lapangan. Hasil evaluasi keseluruhan disajikan pada Tabel V, sedangkan nilai Average Precision (AP) untuk setiap kelas disajikan pada Tabel VI.

Tabel V. Hasil Evaluasi Cross-Domain Model A Dan Model B

Model	Domain Uji	mAP50	mAP50-95	Precision	Recall	F1
Model A	Publik	0,6989	0,4857	0,7709	0,6254	0,6906
Model A	Lapangan	0,4491	0,3121	0,4876	0,4633	0,4751
Model B	Publik	0,7071	0,5076	0,7491	0,6539	0,6983
Model B	Lapangan	0,4056	0,2882	0,5337	0,4611	0,4948

Tabel VI. Average Precision (AP) Per Kelas Pada Masing-Masing Domain Uji

Model	Domain Uji	bercak_daun	daun_sehat	keriting_daun	virus_kuning
Model A	Publik	0,4909	0,7201	0,7875	0,7972
Model A	Lapangan	0,3156	0,5244	0,4897	0,4667
Model B	Publik	0,4982	0,7423	0,7916	0,7964
Model B	Lapangan	0,2999	0,5250	0,3714	0,4262

Berdasarkan Tabel V, kedua model mengalami penurunan performa ketika diuji pada dataset lapangan. Nilai mAP50 Model A menurun dari 0,6989 menjadi 0,4491 (sekitar 35,7%), sedangkan Model B menurun dari 0,7071 menjadi 0,4056 (sekitar 42,6%). Penurunan serupa juga terlihat pada metrik precision, recall, F1-score, dan mAP50-95. Temuan ini menunjukkan bahwa kedua model memiliki kemampuan generalisasi yang lebih rendah ketika diterapkan pada domain yang berbeda dibandingkan domain pelatihannya. Pola tersebut konsisten dengan referensi [7] yang melaporkan bahwa model deteksi penyakit tanaman umumnya mengalami penurunan performa ketika diuji pada citra lapangan akibat adanya perbedaan karakteristik visual antar domain. Temuan lain yang menarik adalah bahwa Model B menunjukkan performa sedikit lebih tinggi dibandingkan Model A pada domain publik, namun mengalami penurunan yang lebih besar ketika diuji pada domain lapangan. Hasil ini mengindikasikan bahwa strategi augmentasi yang digunakan pada penelitian ini mampu meningkatkan performa pada domain yang memiliki karakteristik serupa dengan data pelatihan, tetapi belum mampu meningkatkan kemampuan generalisasi pada domain dengan karakteristik visual yang berbeda. Interpretasi ini sejalan dengan referensi [12] yang menjelaskan bahwa augmentasi berbasis transformasi citra sederhana masih memiliki keterbatasan dalam mengurangi kesenjangan visual antar domain.

Pada level kelas, seluruh kelas mengalami penurunan AP ketika model diuji pada dataset lapangan. Penurunan terbesar terjadi pada kelas *keriting_daun* dan *virus_kuning*, sedangkan kelas *daun_sehat* menunjukkan penurunan yang relatif lebih kecil. Pola ini konsisten dengan hasil analisis karakteristik objek pada Tabel III di mana kedua kelas tersebut memiliki skala objek yang lebih kecil pada dataset lapangan dibandingkan dataset publik. Temuan ini sejalan dengan referensi [8] yang melaporkan bahwa perbedaan skala objek antar domain berkaitan dengan penurunan performa deteksi objek. Kelas *bercak_daun* menunjukkan AP yang rendah pada kedua domain, yang menunjukkan bahwa

tantangan deteksi kelas ini bersifat independen dari domain pengujian, selain itu, gejala pada kelas ini tampak memiliki kemiripan visual dengan beberapa kelas penyakit lainnya, sehingga berpotensi menyulitkan model dalam membedakan antar kelas, temuan ini konsisten dengan referensi [13] yang menjelaskan bahwa tingkat keparahan gejala bercak daun (*Cercospora leaf spot*) yang bervariasi pada kondisi lapangan menyulitkan model deep learning dalam mengenali kelas tersebut secara konsisten.

D. Analisa Pola Kesalahan

Pada subbab ini dilakukan analisis pola kesalahan deteksi pada domain lapangan untuk menjelaskan jenis kesalahan yang mendasari penurunan performa model pada evaluasi lintas domain. Analisis difokuskan pada tiga kategori kesalahan, yaitu false negative (FN), false positive (FP), dan misklasifikasi, berdasarkan pencocokan prediksi terhadap ground truth menggunakan ambang Intersection over Union (IoU) sebesar 0,5. Ringkasan masing-masing kategori kesalahan disajikan pada Tabel VII, VIII, dan IX.

Tabel VII. Ringkasan False Negative Per Kelas (Domain Lapangan)

Model	Kelas	Jumlah FN
Model A	<i>bercak_daun</i>	34
Model A	<i>daun_sehat</i>	10
Model A	<i>keriting_daun</i>	34
Model A	<i>virus_kuning</i>	25
Model B	<i>bercak_daun</i>	37
Model B	<i>daun_sehat</i>	4
Model B	<i>keriting_daun</i>	44
Model B	<i>virus_kuning</i>	29

Berdasarkan Tabel VII, kelas *keriting_daun* dan *virus_kuning* menghasilkan jumlah false negative tertinggi pada kedua model. Analisis terhadap luas ground truth pada kasus false negative menunjukkan bahwa rata-rata luas objek yang gagal terdeteksi (0,0208 untuk *keriting_daun* dan 0,0346 untuk *virus_kuning*) berada di bawah median luas objek masing-masing kelas pada Tabel III. Temuan ini mengindikasikan bahwa objek dengan skala lebih kecil cenderung lebih sulit dideteksi dibandingkan objek berukuran lebih besar. Interpretasi tersebut konsisten dengan referensi [8] yang melaporkan bahwa objek berukuran

kecil pada citra lapangan memiliki peluang lebih tinggi untuk tidak terdeteksi oleh model deteksi objek. Temuan ini juga sejalan dengan referensi [14] yang melaporkan bahwa performa deteksi objek berukuran kecil pada berbagai varian YOLO cenderung lebih rendah pada lingkungan lapangan yang kompleks dibandingkan objek berukuran lebih besar.

20260503_092503_jpg.rf.9a2bb3f4c79aa0e413beae7d82811ae2.jpg



Gambar 4. Contoh kegagalan deteksi (false negative) pada kelompok objek *keriting_daun*

Selain ukuran objek, pengamatan pada Gambar 4 menunjukkan bahwa beberapa kasus false negative terjadi ketika beberapa daun saling berdekatan (crowding) sehingga batas antar objek menjadi kurang jelas. Kondisi ini menyulitkan model dalam memisahkan setiap objek sebagai instance yang berbeda. Pola tersebut konsisten dengan referensi [8] yang menjelaskan bahwa kompleksitas latar belakang dan kepadatan objek dapat menurunkan performa deteksi pada citra lapangan.

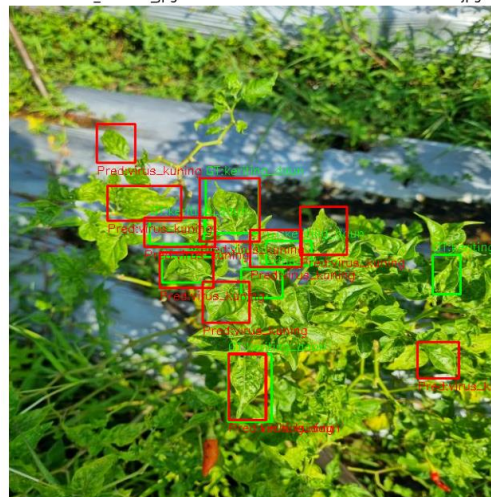
Tabel VIII. Ringkasan False Positive Per Kelas (Domain Lapangan)

Model	Kelas	Jumlah FP
Model A	<i>bercak_daun</i>	18
Model A	<i>daun_sehat</i>	43
Model A	<i>keriting_daun</i>	27
Model A	<i>virus_kuning</i>	66
Model B	<i>bercak_daun</i>	19
Model B	<i>daun_sehat</i>	41
Model B	<i>keriting_daun</i>	44
Model B	<i>virus_kuning</i>	31

Pada Tabel VIII, kelas *virus_kuning* menunjukkan jumlah FP tertinggi pada Model A (66 kasus). Pengamatan visual terhadap sampel FP dengan confidence tertinggi menunjukkan bahwa model mengenali pola belang atau warna

kekuningan pada daun sebagai gejala *virus_kuning*, termasuk pada kasus di mana warna tersebut merupakan variasi pigmentasi alami daun atau efek pencahayaan. Kesalahan ini sejalan dengan penelitian referensi [15] yang melaporkan bahwa perubahan warna daun merupakan salah satu karakteristik visual yang paling dominan pada deteksi penyakit tanaman sehingga berpotensi memicu false positive.

20260503_092416_jpg.rf.0526f0dd5ea719c306859c4d53de7081.jpg

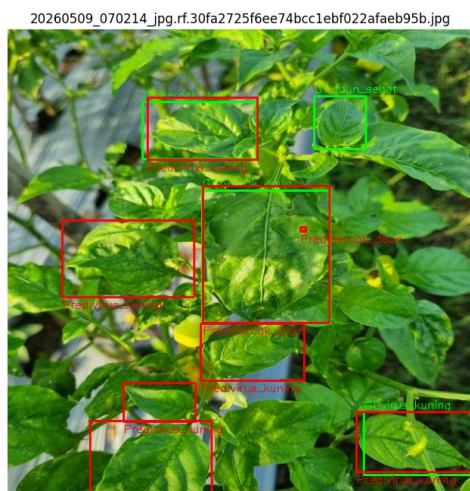


Gambar 5. Contoh false positive kelas *virus_kuning*

Berdasarkan Tabel IX, misklasifikasi paling sering mengarah ke kelas *virus_kuning*, terutama dari kelas *keriting_daun* dan *bercak_daun*. Pola tersebut muncul secara konsisten pada kedua model sehingga menunjukkan bahwa kesalahan klasifikasi tidak hanya dipengaruhi oleh model tertentu, tetapi juga berkaitan dengan karakteristik visual antar kelas pada dataset lapangan.

Tabel IX. Ringkasan Misklasifikasi Antar Kelas (Domain Lapangan)

Model	Ground Truth	Prediksi	Jumlah
Model A	<i>bercak_daun</i>	<i>virus_kuning</i>	5
Model A	<i>keriting_daun</i>	<i>virus_kuning</i>	13
Model A	<i>virus_kuning</i>	<i>bercak_daun / daun_sehat</i>	7
Model B	<i>bercak_daun</i>	<i>virus_kuning</i>	8
Model B	<i>virus_kuning</i>	<i>bercak_daun / daun_sehat / keriting_daun</i>	9



Gambar 6. Contoh misklasifikasi kelas *bercak_daun* menjadi *virus_kuning*

Pengamatan terhadap contoh misklasifikasi pada Gambar 6 menunjukkan bahwa objek yang salah diklasifikasikan memiliki rentang luas bounding box yang bervariasi, termasuk objek berukuran relatif besar. Hal ini mengindikasikan bahwa kesalahan klasifikasi tersebut tidak hanya berkaitan dengan ukuran objek, tetapi juga dengan karakteristik visual lain yang dimiliki objek tersebut. Berdasarkan observasi terhadap contoh hasil deteksi, beberapa gejala *bercak_daun* tampak memiliki pola warna yang menyerupai *virus_kuning*, sehingga model menghasilkan prediksi dengan nilai confidence yang relatif tinggi (0,68–0,93). Interpretasi ini konsisten dengan referensi [15][16] yang melaporkan bahwa kemiripan karakteristik visual antar penyakit tanaman dapat meningkatkan kemungkinan terjadinya misklasifikasi.

E. Sintesis

Bagian ini menyajikan sintesis yang menghubungkan karakteristik dataset, hasil evaluasi performa, dan pola kesalahan ke dalam satu kerangka interpretasi, sebagaimana disajikan pada Tabel X.

Tabel X. Sintesis Temuan Penelitian

Karakteristik Dataset	Hasil Evaluasi	Pola Error	Interpretasi
Bbox <i>keriting_daun</i> & <i>virus_kuning</i> di lapangan jauh lebih kecil (median 0,026 vs	AP kedua kelas turun paling besar di lapangan (<i>keriting_daun</i> 0,79 menjadi 0,49;	FN didominasi median luas kelas; ditemukan kasus crowding	Ukuran objek kecil konsisten dengan penurunan performa & FN tinggi

0,064; 0,042 vs 0,080)	<i>virus_kuning</i> 0,80 menjadi 0,47)	FP tertinggi pada <i>virus_kuning</i> (66 kasus Model A), terkait pola belang/pucat alami	Homogenitas visual tidak menyederhanakan deteksi; warna kekuningan alami tetap memicu FP
Sebaran brightness & sharpness lapangan lebih sempit/homogen vs publik	Precision relatif stabil kedua domain; recall menurun tajam di lapangan		Kemiripan karakteristik visual antar kelas berkaitan dengan meningkatnya misklasifikasi; augmentasi yang digunakan belum meningkatkan generalisasi lintas domain.
Gejala <i>bercak_daun</i> & <i>keriting_daun</i> di lapangan tumpang tindih warna dengan <i>virus_kuning</i>	Model B (augmentasi) tidak memperbaiki generalisasi meski performa publik sedikit lebih baik	Misklasifikasi <i>bercak_daun</i> & <i>keriting_daun</i> paling sering → <i>virus_kuning</i> , konsisten pada kedua model	

Berdasarkan Tabel X, hasil penelitian menunjukkan adanya keterkaitan yang konsisten antara karakteristik visual dataset, performa deteksi, dan pola kesalahan pada domain lapangan. Perbedaan skala objek, khususnya pada kelas *keriting_daun* dan *virus_kuning*, konsisten dengan penurunan Average Precision (AP) yang lebih besar pada kedua kelas tersebut. Analisis false negative memperkuat temuan ini, di mana objek yang gagal terdeteksi umumnya memiliki ukuran di bawah median kelasnya serta sering muncul pada kondisi crowding. Temuan ini sejalan dengan referensi [8][9] yang menjelaskan bahwa perbedaan skala objek antar domain berkaitan dengan menurunnya kemampuan generalisasi model deteksi objek.

Selain skala objek, karakteristik warna juga berkaitan dengan pola kesalahan yang diamati. Meskipun dataset lapangan memiliki distribusi brightness dan sharpness yang lebih homogen, kondisi tersebut tidak mengurangi kesalahan deteksi. False positive pada kelas *virus_kuning* banyak muncul pada daun yang memiliki warna kekuningan atau pucat secara alami, sedangkan misklasifikasi paling sering terjadi dari kelas *bercak_daun* dan *keriting_daun* menuju *virus_kuning*. Berdasarkan pengamatan terhadap contoh hasil deteksi, pola tersebut mengindikasikan bahwa kemiripan karakteristik visual antar kelas berkontribusi terhadap meningkatnya kesalahan klasifikasi, sejalan dengan temuan referensi [15].

Temuan tersebut juga menjelaskan mengapa Model B yang menggunakan augmentasi tidak menunjukkan peningkatan kemampuan generalisasi meskipun memiliki performa sedikit lebih baik pada domain publik. Hasil ini mengindikasikan bahwa augmentasi yang diterapkan pada penelitian ini belum mampu merepresentasikan variasi visual utama yang dijumpai pada dataset lapangan, terutama perbedaan skala objek dan karakteristik visual antar kelas. Interpretasi ini konsisten dengan referensi [12] yang menyatakan bahwa transformasi citra sederhana masih memiliki keterbatasan dalam mengurangi kesenjangan visual antar domain.

Secara keseluruhan, sintesis hasil penelitian menunjukkan bahwa penurunan performa model pada domain lapangan tidak berkaitan dengan satu karakteristik dataset saja, melainkan merupakan hasil interaksi beberapa karakteristik visual yang saling berhubungan. Perbedaan skala objek berkaitan dengan meningkatnya false negative, sedangkan kemiripan karakteristik visual antar kelas berkaitan dengan meningkatnya false positive dan misklasifikasi. Dengan demikian, penelitian ini tidak hanya mengonfirmasi adanya domain shift sebagaimana ditunjukkan pada penelitian sebelumnya [10], tetapi juga memberikan penjelasan yang lebih rinci mengenai bagaimana karakteristik visual dataset tercermin pada pola kesalahan deteksi YOLO.

IV. KESIMPULAN

Penelitian ini menganalisis perbedaan karakteristik visual antara dataset publik dan dataset lapangan serta mengevaluasi bagaimana perbedaan tersebut tercermin pada performa dan pola kesalahan deteksi YOLOv8. Hasil penelitian menunjukkan bahwa dataset lapangan memiliki karakteristik visual dan karakteristik objek yang berbeda dibandingkan dataset publik, terutama pada skala objek kelas *keriting_daun* dan *virus_kuning*. Perbedaan tersebut konsisten dengan penurunan performa model pada evaluasi lintas domain, di mana false negative lebih banyak dijumpai pada objek berukuran kecil dan kondisi crowding, sedangkan false positive serta

misklasifikasi berkaitan dengan kemiripan karakteristik visual antar kelas penyakit. Selain itu, augmentasi standar pada Model B belum menunjukkan peningkatan kemampuan generalisasi pada domain lapangan.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa analisis karakteristik dataset dan pola kesalahan dapat melengkapi evaluasi performa model dalam menjelaskan fenomena *domain shift* pada deteksi penyakit daun cabai. Pendekatan ini diharapkan dapat menjadi acuan dalam pengembangan model yang lebih robust terhadap variasi kondisi lapangan, baik melalui strategi augmentasi, adaptasi domain, maupun pengayaan dataset lapangan. Penelitian selanjutnya disarankan untuk mengembangkan strategi augmentasi dan adaptasi domain yang lebih sesuai dengan karakteristik visual citra lapangan serta memperluas variasi dataset lapangan guna meningkatkan kemampuan generalisasi model pada kondisi implementasi nyata.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Direktorat Riset dan Pengabdian kepada Masyarakat, Direktorat Jenderal Riset dan Pengembangan, Kementerian Pendidikan Tinggi, Sains, dan Teknologi, serta Universitas Duta Bangsa Surakarta atas pendanaan yang diberikan melalui nomor kontrak 285/C3/DT.05.00/PL-BARU/2026, 042/LL6/AL.04.03/PL-BARU/2026, dan 066/UDB.LPPM/A.34-HK/II/2026.

REFERENSI

- [1] E. Umihanni, D. Haryono, K. Murniati, dan T. Endaryanto, "Sustainability of pepper cultivation in Lampung, Indonesia, with rap-pepper coordination techniques and pepper threshing technology," *Edelweiss Applied Science and Technology*, vol. 10, no. 2, pp. 800–811, 2026.
- [2] F. Ahmad, K. Kusumiyati, M. A. Soleh, M. R. Khan, dan R. S. Sundari, "Chili cultivars vulnerability: a multi-factorial examination of disease and pest-induced yield decline across different growing microclimates and watering regimens," *BMC Plant Biol.*, vol. 24, no. 1, p. 979, 2024.
- [3] M. G. F. Odounfa, C. G. Hounmenou, V. K. Salako, A. Affokpon, dan R. L. G. Kakai, "Deep learning enables precision agriculture for sustainable chili pepper disease detection in Benin," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 315, 2025.
- [4] K. R. Kanaparthi dan S. S. Ilango, "Deep learning framework for timely detection and classification of chili leaf diseases and pests," *Sci. Rep.*, vol. 16, no. 1, p. 4360, 2026.
- [5] J. Terven, D.-M. Córdova-Esparza, dan J.-A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer

- Vision: From YOLOv1 to YOLOv8 and YOLO-NAS,” *Mach. Learn. Knowl. Extr.*, vol. 5, no. 4, pp. 1680–1716, 2023.
- [6] A. Upadhyay et al., “Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture,” *Artif. Intell. Rev.*, vol. 58, no. 3, p. 92, 2025.
- [7] X. Wu, X. Fan, P. Luo, S. Das Choudhury, T. Tjahjadi, dan C. Hu, “From Laboratory to Field: Unsupervised Domain Adaptation for Plant Disease Recognition in the Wild,” *Plant Phenomics*, vol. 5, p. 0038, 2023.
- [8] M. Shoaib et al., “An advanced deep learning models-based plant disease detection: A review of recent research,” *Front. Plant Sci.*, vol. 14, 2023.
- [9] S. Yang, Q. Feng, J. Zhang, W. Yang, W. Zhou, dan W. Yan, “From laboratory to field: cross-domain few-shot learning for crop disease identification in the field,” *Front. Plant Sci.*, vol. 15, 2024.
- [10] S. Zhu, P. Zhu, Y. Wu, dan W. Qiao, “Cross-Domain Object Detection with Hierarchical Multi-Scale Domain Adaptive YOLO,” *Sensors*, vol. 25, no. 17, p. 5363, 2025.
- [11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, dan A. Zisserman, “The Pascal Visual Object Classes (VOC) Challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [12] M. Xu, S. Yoon, A. Fuentes, J. Yang, dan D. S. Park, “Style-Consistent Image Translation: A Novel Data Augmentation Paradigm to Improve Plant Disease Recognition,” *Front. Plant Sci.*, vol. 12, 2022.
- [13] D. V. Leite, A. V. de Brito, G. G. Faccioli, dan G. Haddad Souza Vieira, “Deep Learning Models for Detection and Severity Assessment of Cercospora Leaf Spot (*Cercospora capsici*) in Chili Peppers Under Natural Conditions,” *Plants*, vol. 14, no. 13, p. 2011, 2025.
- [14] R. Sapkota, Z. Meng, M. Churuvija, X. Du, Z. Ma, dan M. Karkee, “Comprehensive performance evaluation of YOLOv12, YOLO11, YOLOv10, YOLOv9 and YOLOv8 on detecting and counting fruitlet in complex orchard environments,” *Agriculture Communications*, vol. 4, no. 1, p. 100125, 2026.
- [15] T. Jiang dan T. Zhou, “Unraveling the Mechanisms of Virus-Induced Symptom Development in Plants,” *Plants*, vol. 12, no. 15, p. 2830, 2023.
- [16] M. S. Krishna, P. Machado, R. I. Otuka, S. W. Yahaya, F. Neves dos Santos, dan I. K. Ihianle, “Plant Leaf Disease Detection Using Deep Learning: A Multi-Dataset Approach,” *J (Basel)*, vol. 8, no. 1, p. 4, 2025.