

Fusi Fitur Deep dan Manual Berbasis Masker Warna dengan Seleksi F-ANOVA Klasifikasi Penyakit Padi

Agung Ramadhanu^{1✉}

¹Magister Teknik Informatika, Fakultas Ilmu Komputer, Universitas Putra Indonesia YPTK, Jl. Raya Lubuk Begalung, 25221, Padang, Indonesia

✉ agung_ramadhanu@upiyptk.ac.id

Abstract

Studies on rice leaf disease classification frequently report accuracy above 99%, yet such figures can be inflated when duplicate images are shared between training and test partitions. This study evaluates a dual-branch hybrid pipeline for rice leaf disease classification under a leakage-controlled evaluation protocol. A public Kaggle dataset of 7,421 images was deduplicated by exact perceptual-hash matching, leaving 4,410 unique images in five classes. Deep features (1,280 dimensions) from an ImageNet-pretrained EfficientNet-B0 were combined with 73 handcrafted descriptors computed from colour-segmentation masks (colour moments, masked GLCM, LBP, and mask morphometry). The proposed scenario (E5) keeps the 128 features with the highest ANOVA F-scores, with scaling and selection fitted on training folds only. Five ablation scenarios (E1–E5) were evaluated with SVM-RBF and LightGBM using repeated stratified 5-fold cross-validation with three seeds (15 test folds) and Wilcoxon signed-rank tests. With SVM-RBF, E5 reached $99.64\% \pm 0.18\%$ accuracy and $99.47\% \pm 0.28\%$ macro-F1 while using 90.5% fewer features than naive concatenation (1,353 features). E5 was better than handcrafted features alone ($p = 0.0067$; Holm-adjusted $p = 0.034$), but this was not significant under a resampled t-test corrected for fold dependence ($p = 0.17$), and E5 was not significantly different from deep features alone or naive concatenation ($p > 0.05$). With LightGBM, E5 was significantly worse than E1, E3, and E4. Visual inspection of one image per class showed that the colour-based masks do not reliably isolate visible lesions, so the handcrafted features should not be read as lesion-specific. The compact representation is therefore competitive rather than superior, and its behaviour depends on the classifier. All results are in-distribution; field generalisation was not tested.

Keywords: Rice Leaf Disease Classification, Feature Fusion, EfficientNet-B0, Data Leakage, Ablation Study

Abstrak

Penelitian klasifikasi penyakit daun padi sering melaporkan akurasi di atas 99%, tetapi angka tersebut dapat menggelembung bila citra duplikat terbagi ke data latih dan data uji. Penelitian ini mengevaluasi *pipeline* hibrida dua cabang untuk klasifikasi penyakit daun padi dengan protokol evaluasi yang mengendalikan kebocoran data. Dataset publik Kaggle sebanyak 7.421 citra dideduplikasi berdasarkan kecocokan *perceptual hash* identik sehingga menghasilkan 4.410 citra unik pada lima kelas. Fitur *deep* (1.280 dimensi) dari EfficientNet-B0 berbobot ImageNet digabungkan dengan 73 deskriptor buatan tangan dari masker segmentasi warna (momen warna, GLCM bertopeng, LBP, dan morfometri masker). Skenario usulan (E5) mempertahankan 128 fitur dengan skor F-ANOVA tertinggi; standardisasi dan seleksi dihitung hanya pada data latih tiap lipatan. Lima skenario ablasi (E1–E5) dievaluasi dengan SVM-RBF dan LightGBM memakai validasi silang *stratified 5-lipatan* berulang dengan tiga *seed* (15 lipatan uji) serta uji peringkat bertanda Wilcoxon. Dengan SVM-RBF, E5 mencapai akurasi $99,64\% \pm 0,18\%$ dan Macro-F1 $99,47\% \pm 0,28\%$ dengan fitur 90,5% lebih sedikit daripada penggabungan naif (1.353 fitur). E5 lebih baik daripada fitur buatan tangan saja ($p = 0,0067$; terkoreksi Holm 0,034), tetapi tidak signifikan pada uji-t terkoreksi ketergantungan lipatan ($p = 0,17$), dan tidak berbeda signifikan dari fitur *deep* saja maupun penggabungan naif ($p > 0,05$). Dengan LightGBM, E5 signifikan lebih buruk daripada tiga dari empat pembandingan. Representasi ringkas ini dengan demikian kompetitif, bukan unggul, dan perilakunya bergantung pada pengklasifikasi. Seluruh hasil bersifat *in-distribution*; generalisasi ke citra lapang tidak diuji.

Kata kunci: Klasifikasi Penyakit Daun Padi, Fusi Fitur, EfficientNet-B0, Kebocoran Data, Kajian Ablasi

KomtekInfo is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.

1. Pendahuluan

Penyakit daun padi, antara lain hawar daun bakteri dan blas, dilaporkan berdampak besar terhadap kualitas dan kuantitas produksi padi [1]. Pengenalan penyakit dari citra daun karena itu menjadi topik aktif dalam visi komputer pertanian. Pendekatan yang umum adalah mengekstraksi fitur dari jaringan konvolusional (CNN) lalu mengklasifikasikannya dengan pengklasifikasi lain. Salah satu contohnya adalah kerangka hibrida yang memadukan MobileNetV3 dengan pengklasifikasi CatBoost untuk penyakit daun padi [2]. Fitur dari beberapa jaringan pralatih juga dapat digabungkan langsung lalu dimasukkan ke jaringan saraf [3].

Penggabungan fitur *deep* dengan deskriptor buatan tangan juga telah dilaporkan untuk penyakit padi. Uddin dan Joolee [4] memadukan fitur InceptionResNetV2

dengan deskriptor tekstur LNEP pada dataset Paddy Doctor, dan Aggarwal dkk. [5] mengekstraksi fitur dari 32 model pralatih untuk tiga penyakit padi pada 551 citra, dengan akurasi tertinggi 94% pada citra tersegmentasi dibanding 91% pada citra tanpa segmentasi. Penelitian ini tidak berusaha mengungguli angka-angka tersebut, yang diperoleh pada dataset dan protokol berbeda. Fokusnya adalah seberapa jauh akurasi bertahan bila kebocoran data dikendalikan, dan apakah selisih antar-skenario bermakna secara statistik.

Klaim akurasi yang sangat tinggi pada dataset publik perlu diinterpretasikan secara kritis. Tinjauan terhadap studi penyakit tomat menunjukkan bahwa hasil pada dataset berlatar seragam cenderung terlalu optimistis [6]. Pada PlantVillage, model yang hanya melihat delapan piksel latar belakang masih mencapai akurasi 49,0% pada 38 kelas, jauh di atas peluang acak 2,6% [7]. Kebocoran

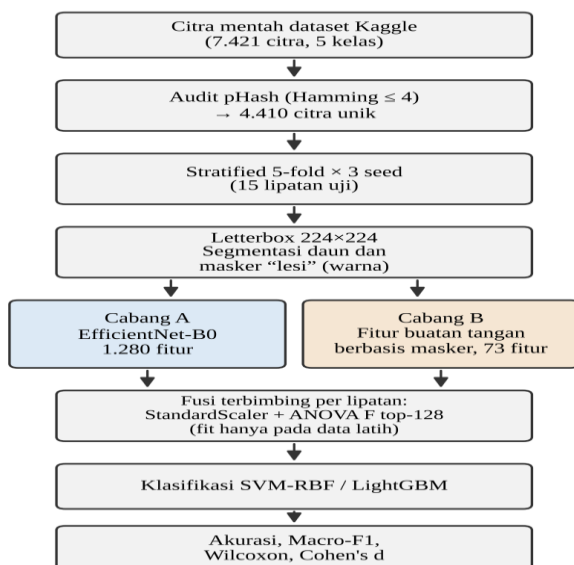
data, yaitu informasi dari data uji yang ikut memengaruhi pelatihan, juga terdokumentasi luas sebagai penyebab hasil yang terlalu optimistis [8]. Salah satu sumbernya pada dataset citra adalah duplikat atau citra hasil augmentasi yang jatuh ke data latih dan data uji sekaligus. Pada CIFAR-10 dan CIFAR-100, misalnya, sekitar 3,3% dan 10% citra uji memiliki *near-duplicate* di data latih [9]. Pada dataset padi yang dipakai dalam penelitian ini, audit awal menemukan bahwa 3.011 dari 7.421 citra (40,6%) memiliki nilai *perceptual hash* yang sama dengan citra lain dan karenanya dikeluarkan sebagai duplikat.

Dua masalah metodologis lain relevan di sini. Pertama, penggabungan langsung fitur *deep* dan fitur buatan tangan menghasilkan vektor berdimensi besar (pada penelitian ini 1.353 fitur) yang belum tentu semuanya informatif. Seleksi fitur pun dapat menjadi sumber bias: bila dilakukan pada seluruh data sebelum validasi silang, galat yang diestimasi menjadi terlalu optimistis [10]. Kedua, evaluasi *in-distribution* tidak menjamin kinerja di lapang. Model yang dilatih pada citra laboratorium dilaporkan tidak mencapai akurasi yang sama pada citra lapang, dan penghilangan latar belakang membantu generalisasi pada kasus jagung [11]. Penelitian ini tidak menguji generalisasi lapang; fokusnya adalah evaluasi *in-distribution* yang dikendalikan kebocorannya.

Penelitian ini bertanya: (1) apakah seleksi 128 fitur secara terbimbing mempertahankan kinerja dibanding penggabungan naif 1.353 fitur, dan (2) apakah fitur buatan tangan berbasis masker memberi informasi tambahan terhadap fitur *deep*. Kontribusinya: (a) kurasi dataset dengan *perceptual hashing* sebelum pembagian data; (b) protokol evaluasi yang menghitung standarisasi dan seleksi fitur hanya pada data latih tiap lipatan; dan (c) kajian ablasi lima skenario dengan uji Wilcoxon yang melaporkan hasil apa adanya, termasuk skenario yang tidak menguntungkan metode usulan.

2. Metodologi Penelitian

Alur penelitian ditunjukkan pada Gambar 1. Seluruh tahap setelah pembagian data dijalankan pada tiap lipatan secara terpisah.



Gambar 1. Alur Penelitian

2.1 Dataset dan Kurasi

Dataset primer adalah *Rice Leaf Disease Detection Dataset* dari Kaggle [12] berisi 7.421 citra. Setiap citra diberi *perceptual hash* (pHash) 64-bit berbasis transformasi kosinus diskret. Citra dengan nilai pHash identik dianggap duplikat, dan hanya citra pertama pada urutan berkas yang dipertahankan. Proses ini menyisakan 4.410 citra unik pada lima kelas (Tabel 1). Kriteria ini tidak menangkap citra yang mirip tetapi berhash berbeda (Bagian 3.3). Deduplikasi dilakukan sebelum pembagian data sehingga duplikat tidak terbagi ke data latih dan data uji.

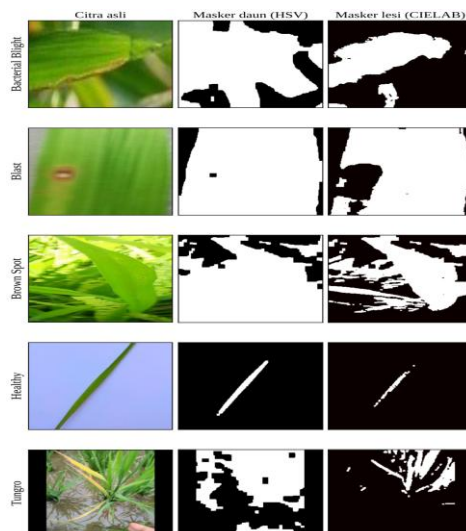
Tabel 1. Distribusi Kelas Setelah Kurasi

Kelas	Jumlah citra	Proporsi (%)
Healthy	1.483	33,63
Tungro	1.060	24,04
Brownspot	757	17,17
Bacterialblight	569	12,90
Blast	541	12,27
Total	4.410	100,00

2.2 Praproses dan Segmentasi

Citra diubah ke ukuran 224×224 piksel dengan *letterbox* (menjaga rasio aspek, *padding* abu-abu netral) agar bentuk objek memanjang tidak terdistorsi. Segmentasi daun memakai ruang warna HSV: kanal saturasi dipisahkan dengan ambang Otsu, lalu dirapikan dengan operasi morfologi *opening* (kernel 5×5) dan *closing* (kernel 9×9). Segmentasi "lesi" mengelompokkan piksel daun pada kanal a^*b^* ruang warna CIELAB dengan K-Means ($k = 3$); klaster yang dipilih berdasarkan kroma, dirancang untuk warna cokelat, kuning, atau nekrotik, dipakai sebagai masker lesi.

Gambar 2 menampilkan hasil segmentasi pada citra pertama tiap kelas. Pemeriksaan visual ini menunjukkan keterbatasan. Pada citra berlatar kompleks (*Brown Spot*, *Tungro*) masker daun mencakup banyak area non-daun. Pada citra *Bacterial Blight* dan *Blast*, masker "lesi" mencakup sebagian besar area daun hijau, sedangkan bercak cokelat yang tampak pada citra *Blast* justru berada di luar masker. Segmentasi tidak divalidasi terhadap anotasi, sehingga kualitas kuantitatifnya tidak diketahui.



Gambar 2. Segmentasi pada Citra Pertama Tiap Kelas

2.3 Ekstraksi Fitur Dua Cabang

Cabang A (deep). EfficientNet-B0 [13] berbobot ImageNet dipakai sebagai pengekstraksi fitur. Keluaran *global average pooling* sebanyak 1.280 dimensi diambil dari citra daun tersegmentasi (atau citra mentah pada skenario tertentu). Ekstraksi berjalan pada GPU dengan presisi campuran FP16.

Cabang B (fitur buatan tangan berbasis masker, 73 fitur). Terdiri atas 18 fitur momen warna (rerata, simpangan baku, *skewness* pada HSV dan CIELAB), 40 fitur GLCM [14] bertopeng (jarak 1 dan 3, empat orientasi, lima properti: kontras, disimilaritas, homogenitas, energi, korelasi), 10 fitur histogram LBP seragam ($P = 8$, $R = 1$), dan 5 fitur morfometri masker (rasio luas masker terhadap daun, *compactness*, *solidity*, eksentrisitas, dan jumlah komponen terhubung). GLCM dihitung hanya pada piksel di dalam masker sehingga batas latar tidak ikut dihitung. Karena masker tidak selalu mengisolasi lesi (Gambar 2), fitur ini tidak diasumsikan spesifik lesi.

2.4 Fusi Fitur Terbimbing

Pada tiap lipatan, fitur *deep* dan fitur buatan tangan distandardisasi terpisah (*StandardScaler* dihitung pada data latih), digabung, lalu diurutkan dengan skor F-ANOVA [15]:

$$F = \text{MSB} / \text{MSW} \quad (1)$$

dengan MSB rerata kuadrat antar-kelas dan MSW rerata kuadrat dalam-kelas. Sebanyak $K = 128$ fitur dengan skor tertinggi dipertahankan. Rerata, simpangan baku, dan indeks fitur terpilih diperoleh hanya dari data latih, lalu diterapkan apa adanya pada data uji. Seleksi di dalam lipatan diperlukan karena seleksi pada seluruh data sebelum validasi silang menaikkan estimasi kinerja secara bias [10].

2.5 Skenario Ablasi dan Pengklasifikasi

Lima skenario dibandingkan (Tabel 2). Perlu dicatat bahwa E4 dan E5 berbeda pada dua hal sekaligus, yaitu masukan citra (mentah vs tersegmentasi) dan adanya seleksi fitur, sehingga selisih keduanya tidak dapat diatribusikan pada seleksi saja.

Tabel 2. Skenario Ablasi

Kode	Masukan fitur	Dimensi
E1	<i>Deep</i> , citra mentah	1.280
E2	<i>Deep</i> , citra daun tersegmentasi	1.280
E3	Fitur buatan tangan berbasis masker	73
E4	<i>Deep</i> (mentah) + manual (tanpa topeng), digabung tanpa seleksi	1.353
E5	<i>Deep</i> + manual (tersegmentasi), seleksi F-ANOVA	128

Pengklasifikasi yang dipakai adalah SVM-RBF [16] dengan $C = 5,0$ dan LightGBM [17] dengan 60 estimator, *learning rate* 0,08, dan 15 daun. Parameter lain memakai nilai bawaan pustaka dan tidak dituning pada data uji.

2.6 Protokol Evaluasi dan Uji Statistik

Evaluasi memakai validasi silang *stratified* 5-lipatan yang diulang pada tiga *seed* (42, 123, 2024), sehingga diperoleh 15 lipatan uji per skenario dan pengklasifikasi. Karena hiperparameter ditetapkan tetap, validasi silang ini

tidak bersarang (*nested*). Metrik yang dilaporkan adalah akurasi dan Macro-F1:

$$\text{Macro-F1} = (1/C) \sum_c 2P_c R_c / (P_c + R_c) \quad (2)$$

Perbandingan E5 terhadap E1–E4 diuji dengan uji peringkat bertanda Wilcoxon [18] pada 15 pasang nilai Macro-F1 untuk tiap pengklasifikasi, dengan $\alpha = 0,05$. Ukuran efek dihitung sebagai Cohen's d berpasangan:

$$d = \bar{D} / s_D \quad (3)$$

dengan $\bar{D}$ dan s_D masing-masing rerata dan simpangan baku selisih Macro-F1 berpasangan. Nilai p Wilcoxon dilaporkan apa adanya, lalu dua analisis sensitivitas ditambahkan: koreksi Holm [19] atas delapan perbandingan, dan uji- t berpasangan terkoreksi Nadeau–Bengio [20], [21] dengan varians selisih dikali $(1/15 + 1/4)$, karena ke-15 lipatan tidak saling bebas. Uji terkoreksi ini adalah pendekatan dan tidak divalidasi khusus untuk skema berulang dengan beberapa *seed*. Implementasi memakai Python 3.12.6, PyTorch, scikit-learn, dan LightGBM; ekstraksi fitur *deep* dijalankan pada GPU NVIDIA GeForce RTX 3050.

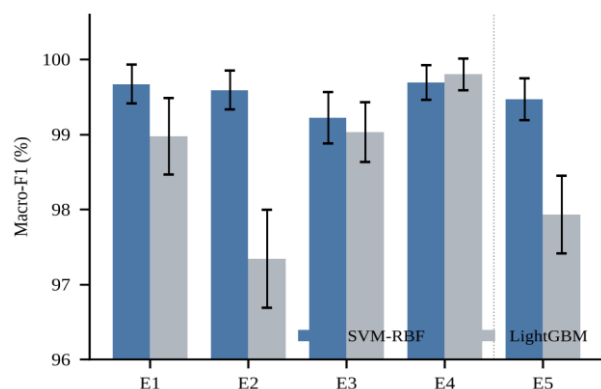
3. Hasil dan Pembahasan

3.1 Kinerja Skenario Ablasi

Tabel 2 dan Gambar 3 merangkum hasil pada 15 lipatan uji. Dengan SVM-RBF, seluruh skenario mencapai Macro-F1 di atas 99%. Skenario usulan E5 mencapai akurasi $99,64\% \pm 0,18\%$ dan Macro-F1 $99,47\% \pm 0,28\%$. Nilai ini secara numerik lebih rendah daripada E1 (99,67%), E2 (99,59%), dan E4 (99,69%), dan lebih tinggi daripada E3 (99,22%). Selisihnya dengan E1, E2, dan E4 berkisar 0,11–0,22 poin persentase, lebih kecil daripada simpangan baku antar-lipatan (0,23–0,28).

Tabel 3. Hasil Ablasi (rerata $\pm$ SD, 15 lipatan uji, %)

Skenario	Fitur	Model	Akurasi	Macro-F1
E1	1.280	SVM-RBF	99,76 $\pm$ 0,18	99,67 $\pm$ 0,26
		LightGBM	99,29 $\pm$ 0,34	98,97 $\pm$ 0,51
E2	1.280	SVM-RBF	99,71 $\pm$ 0,18	99,59 $\pm$ 0,26
		LightGBM	98,16 $\pm$ 0,45	97,34 $\pm$ 0,65
E3	73	SVM-RBF	99,42 $\pm$ 0,23	99,22 $\pm$ 0,34
		LightGBM	99,18 $\pm$ 0,36	99,03 $\pm$ 0,40
E4	1.353	SVM-RBF	99,78 $\pm$ 0,16	99,69 $\pm$ 0,23
		LightGBM	99,86 $\pm$ 0,15	99,80 $\pm$ 0,21
E5	128	SVM-RBF	99,64 $\pm$ 0,18	99,47 $\pm$ 0,28
		LightGBM	98,54 $\pm$ 0,36	97,93 $\pm$ 0,52



Gambar 3. Macro-F1 per skenario (rerata $\pm$ SD). Sumbu-y dipotong pada 96% agar selisih terlihat.

SVM-RBF memberi Macro-F1 lebih tinggi daripada LightGBM pada E1, E2, E3, dan E5, tetapi tidak pada E4: LightGBM pada penggabungan naif 1.353 fitur menghasilkan nilai tertinggi di seluruh tabel (Macro-F1 99,80%). Pada LightGBM, E5 (97,93%) lebih rendah daripada E1, E3, dan E4, dengan selisih terbesar terhadap E4.

3.2 Signifikansi Statistik

Tabel 4. Uji Statistik E5 terhadap Pembandingan

Banding	$\Delta F1$ (pp)	W	p	p_{Holm}	p_{NB}	d
SVM: E5-E1	-0,19	32,00	0,1205	0,1912	0,3825	-0,51
SVM: E5-E2	-0,11	27,00	0,0637	0,1912	0,4130	-0,47
SVM: E5-E3	+0,25	14,00	0,0067	0,0336	0,1677	+0,82
SVM: E5-E4	-0,22	28,00	0,0730	0,1912	0,3002	-0,61
LGB: E5-E1	-1,04	2,00	$1,83 \times 10^{-4}$	0,0013	0,0117	-1,63
LGB: E5-E2	+0,59	18,00	0,0151	0,0603	0,2084	+0,74
LGB: E5-E3	-1,10	2,00	$1,83 \times 10^{-4}$	0,0013	0,0096	-1,69
LGB: E5-E4	-1,87	0,00	$6,10 \times 10^{-5}$	$4,9 \times 10^{-4}$	$1,4 \times 10^{-5}$	-3,65

Keterangan: p = Wilcoxon; p_{Holm} = terkoreksi Holm (8 uji); p_{NB} = uji- t terkoreksi Nadeau-Bengio.

Dengan SVM-RBF, hanya perbandingan E5-E3 yang signifikan pada uji Wilcoxon ($p = 0,0067$; $d = +0,82$, efek besar): menambahkan fitur *deep* terseleksi pada fitur buatan tangan menaikkan Macro-F1 sebesar 0,25 poin persentase. Temuan ini bertahan setelah koreksi Holm (0,0336) tetapi tidak pada uji terkoreksi ketergantungan lipatan ($p_{NB} = 0,1677$). Perbandingan E5 dengan E1, E2, dan E4 tidak signifikan pada ketiga prosedur. Hasil ini berarti tidak ditemukan bukti perbedaan, bukan bukti bahwa kinerjanya setara; dengan hanya 15 pasang, uji ini memiliki daya terbatas untuk selisih sekecil 0,2 poin persentase. Dengan LightGBM, E5 lebih buruk daripada E1, E3, dan E4 pada ketiga prosedur ($p_{Holm} \leq 0,0013$; $p_{NB} \leq 0,0117$). Keunggulan E5 atas E2 hanya tampak pada uji Wilcoxon ($p = 0,0151$) dan tidak setelah koreksi Holm (0,0603) maupun uji terkoreksi (0,2084).

3.3 Pembahasan

Temuan utama adalah bahwa 128 fitur terseleksi mempertahankan Macro-F1 di atas 99% pada SVM-RBF dengan dimensi 90,5% lebih kecil daripada E4 (1.353 fitur) dan 90,0% lebih kecil daripada E1 dan E2 (1.280 fitur). Itulah klaim yang didukung data. Data tidak mendukung klaim bahwa E5 lebih akurat daripada fitur *deep* saja atau penggabungan naif. Klaim bahwa representasi yang lebih ringkas lebih tahan terhadap korelasi palsu latar belakang juga tidak diuji dalam penelitian ini. Hal yang sama berlaku untuk manfaat komputasi: dimensi lebih kecil hanya memungkinkan inferensi yang lebih ringan, sedangkan waktu inferensi tidak diukur di sini.

Hasil LightGBM menunjukkan bahwa manfaat seleksi bergantung pada pengklasifikasi. Dugaan penulis, yang belum diuji, adalah bahwa F-ANOVA menilai fitur satu per satu sehingga dapat membuang fitur yang baru informatif lewat interaksi, padahal pohon keputusan memanfaatkan interaksi semacam itu; SVM-RBF yang bekerja pada jarak di ruang fitur lebih toleran terhadap pembuangan tersebut. Pengamatan pada Tabel 3 sejalan dengan dugaan ini tetapi tidak membuktikannya: LightGBM pada 73 fitur buatan tangan utuh (E3)

mencapai Macro-F1 99,03%, sedangkan pada 128 fitur terseleksi (E5) hanya 97,93%, sehingga penurunan tampak terkait pemilihan fitur dan bukan sekadar dimensi yang kecil. Pengujian dugaan ini memerlukan pembandingan seleksi lain, misalnya informasi mutual.

Hasil E3 perlu dibaca hati-hati. Fitur buatan tangan saja mencapai Macro-F1 99,22% dengan SVM-RBF. Karena masker tidak selalu mengisolasi lesi (Gambar 2), angka ini tidak dapat ditafsirkan sebagai bukti bahwa fitur lesi diskriminatif. Fitur tersebut juga menangkap warna dan tekstur daun dan, pada beberapa citra, latar belakang, yang dapat berbeda secara sistematis antar-kelas dalam dataset ini [7]. Pembuktian peran lesi memerlukan masker yang divalidasi terhadap anotasi.

Keterbatasan penelitian ini adalah sebagai berikut. (1) Seluruh evaluasi bersifat *in-distribution* pada satu dataset; skor di atas 99% pada seluruh skenario SVM-RBF menunjukkan dataset ini relatif mudah dan tidak membuktikan kinerja pada citra lapang [6], [11]. (2) Deduplikasi hanya menyingkirkan citra dengan pHash identik. Citra yang mirip tetapi berhash berbeda, serta citra dari daun atau tanaman yang sama, tetap dapat terbagi ke data latih dan uji, dan ambang kemiripan yang lebih longgar tidak diuji. Protokol ini mengendalikan kebocoran, bukan menghapusnya sepenuhnya. (3) Ke-15 lipatan berasal dari tiga pengulangan atas data yang sama, sehingga tidak saling bebas dan nilai p Wilcoxon cenderung terlalu kecil; uji terkoreksi Nadeau-Bengio memberi $p = 0,1677$ untuk E5-E3 dengan SVM-RBF. (4) Delapan uji dijalankan; setelah koreksi Holm p untuk E5-E3 menjadi 0,0336 (tetap di bawah 0,05), tetapi dengan koreksi Bonferroni atas delapan uji menjadi 0,0537 (tidak lagi signifikan). (5) Hiperparameter ditetapkan tetap dan tidak dituning. (6) Desain ablasi tidak memisahkan efek segmentasi dari efek seleksi pada E4 vs E5. (7) Hasil per kelas dan matriks konfusi tidak dianalisis. (8) Segmentasi daun dan lesi tidak divalidasi terhadap anotasi dan terlihat tidak andal pada beberapa citra contoh (Gambar 2).

4. Kesimpulan

Pipeline hibrida dua cabang dengan seleksi 128 fitur secara terbimbing pada protokol terkontrol kebocoran data mencapai akurasi $99,64\% \pm 0,18\%$ dan Macro-F1 $99,47\% \pm 0,28\%$ dengan SVM-RBF, memakai 90,5% lebih sedikit fitur daripada penggabungan naif. Kinerja ini lebih baik daripada fitur buatan tangan saja pada uji Wilcoxon ($p = 0,0067$), tetapi tidak signifikan pada uji terkoreksi ketergantungan lipatan, dan tidak berbeda signifikan dari fitur *deep* saja maupun penggabungan naif. Dengan LightGBM, E5 lebih buruk daripada E1, E3, dan E4. Segmentasi yang dipakai tidak terbukti mengisolasi lesi. Metode usulan karenanya kompetitif dan ringkas, tetapi tidak terbukti lebih unggul, dan hasil ini hanya berlaku pada evaluasi *in-distribution*.

Penelitian lanjutan sebaiknya (1) menilai apakah peta atensi model jatuh pada jaringan yang relevan [22], (2) menguji generalisasi pada citra lapang independen [23], (3) memvalidasi segmentasi terhadap anotasi, membandingkan metode seleksi lain, dan memisahkan efek segmentasi dari seleksi, (4) menganalisis kinerja per kelas, dan (5) mengukur waktu inferensi pada perangkat sasaran.

Daftar Rujukan

- [1] S. Lamba, V. Kukreja, J. Rashid, T. R. Gadekallu, J. Kim, A. Baliyan, D. Gupta, and S. Saini, "A novel fine-tuned deep-learning-based multi-class classifier for severity of paddy leaf diseases," *Front. Plant Sci.*, vol. 14, Art. no. 1234067, 2023, doi: 10.3389/fpls.2023.1234067.
- [2] C. Subbarayudu and M. Kubendiran, "An automated hybrid deep learning framework for paddy leaf disease identification and classification," *Sci. Rep.*, vol. 15, Art. no. 26873, 2025, doi: 10.1038/s41598-025-08071-6.
- [3] M. H. Ishaq and C. Direkoğlu, "A deep learning based fusion model for plant leaf disease classification," in *Proc. 7th Int. Congr. Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*, Ankara, Türkiye, May 2025, doi: 10.1109/ICHORA65333.2025.11017095.
- [4] M. A. Uddin and J. B. Joolee, "Fusion of deep learned and handcrafted features for paddy disease recognition," in *Proc. 2024 Int. Conf. Artif. Intell., Metaverse Cybersecur. (ICAMAC)*, Dubai, UAE, Oct. 2024, Art. no. 10829316, doi: 10.1109/ICAMAC62387.2024.10829316.
- [5] M. Aggarwal, V. Khullar, N. Goyal, A. Singh, A. Tolba, E. B. Thompson, and S. Kumar, "Pre-trained deep neural network-based features selection supported machine learning for rice leaf disease classification," *Agriculture*, vol. 13, no. 5, Art. no. 936, 2023, doi: 10.3390/agriculture13050936.
- [6] M. Jelali, "Deep learning networks-based tomato disease and pest detection: a first review of research studies using real field datasets," *Front. Plant Sci.*, vol. 15, Art. no. 1493322, 2024, doi: 10.3389/fpls.2024.1493322.
- [7] M. A. Noyan, "Uncovering bias in the PlantVillage dataset," arXiv:2206.04374, 2022.
- [8] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
- [9] B. Barz and J. Denzler, "Do we train on test data? Purging CIFAR of near-duplicates," *J. Imaging*, vol. 6, no. 6, Art. no. 41, 2020, doi: 10.3390/jimaging6060041.
- [10] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, no. 10, pp. 6562–6566, 2002, doi: 10.1073/pnas.102102699.
- [11] A. Ahmad, A. El Gamal, and D. Saraswat, "Toward generalization of deep learning-based plant disease identification under controlled and field conditions," *IEEE Access*, vol. 11, pp. 9042–9057, 2023, doi: 10.1109/ACCESS.2023.3240100.
- [12] "Rice Leaf Disease Detection Dataset," Kaggle, pengguna shams069. [Online]. Available: <https://www.kaggle.com/datasets/shams069/rice-leaf-disease-detection-dataset>
- [13] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR 97, 2019, pp. 6105–6114.
- [14] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, 1973, doi: 10.1109/TSMC.1973.4309314.
- [15] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [16] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
- [17] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 3146–3154.
- [18] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945, doi: 10.2307/3001968.
- [19] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.
- [20] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, 2003, doi: 10.1023/A:1024068626366.
- [21] R. R. Bouckaert and E. Frank, "Evaluating the replicability of significance tests for comparing learning algorithms," in *Advances in Knowledge Discovery and Data Mining (PAKDD)*, Lecture Notes in Computer Science, 2004, pp. 3–12, doi: 10.1007/978-3-540-24775-3_3.
- [22] R. Maryani and [Nama Dosen Pembimbing], "Ilusi keterjelasan pada kuantifikasi penajajaran Grad-CAM: studi kasus masker segmentasi warna otomatis pada citra penyakit padi," [jurnal dan status: diisi penulis], 2026.
- [23] R. Maryani and [Nama Dosen Pembimbing], "Evaluasi pergeseran domain zero-shot dan latensi tahap klasifikasi pada pipeline hibrida fitur terseleksi untuk penyakit daun padi," [jurnal dan status: diisi penulis], 2026.