

Klasifikasi Hierarkis dengan Kalibrasi Probabilitas untuk Penyakit Daun Singkong Menggunakan Fusi Fitur dan Random Forest

Hierarchical Classification with Probability Calibration for Cassava Leaf Disease Using Feature Fusion and Random Forest

Regi Fernando Najoan*, Enrico Samuel Djimesha, Argha Orion Silitonga, Wilsen Grivin Mokodaser, George Morris William Tangka

Fakultas Ilmu Komputer, Informatika, Universitas Klabat, Minahasa Utara, Indonesia

Jl. Arnold Mononutu, Airmadidi Atas, Kab. Minahasa Utara, Sulawesi Utara, Indonesia

Email: ^{1*}reginajoan@unklab.ac.id, ²enrico.djimesha@unklab.ac.id, ³argha@unklab.ac.id, ⁴wilsenm@unklab.ac.id, ⁵gtangka@unklab.ac.id

Email Penulis Korespondensi: reginajoan@unklab.ac.id

Received: 27/09/2026, Revised: 11/10/2026, Accepted: 13/10/2026, Published: 18/10/2026

Abstrak—Klasifikasi penyakit daun singkong pada citra lapangan memerlukan probabilitas yang dapat dipercaya dan mekanisme untuk menahan keputusan yang meragukan, bukan hanya akurasi label. Penelitian ini mengevaluasi fusi fitur HSV, GLCM, dan embedding MobileNetV3-Small pada Random Forest hierarkis dua tahap, disertai kalibrasi probabilitas dan prediksi selektif. Sebanyak 7.508 citra CCMT dievaluasi dengan outer stratified 5-fold cross-validation. Hasil perbandingan model dilaporkan sebagai rerata $\pm$ standar deviasi lima outer fold. MobileNetV3-Small end-to-end mencapai accuracy $93,90 \pm 0,73\%$, balanced accuracy $94,64 \pm 0,59\%$, dan Macro F1 $94,23 \pm 0,73\%$, lebih tinggi daripada Hierarchical Hybrid Random Forest dengan $88,28 \pm 1,12\%$, $89,66 \pm 0,76\%$, dan $88,79 \pm 1,07\%$. Selisih Macro F1 hierarkis terhadap Flat Hybrid sebesar 1,03 poin persentase tidak signifikan pada taraf 5% (Wilcoxon dua sisi, $p=0,125$). Evaluasi terpisah untuk kalibrasi dan prediksi selektif membagi setiap outer-train fold menjadi model-fit 70%, calibrator-fit 15%, dan threshold-selection 15%; seluruh partisi tersebut terpisah dari outer-test. Simplex-temperature menghasilkan ECE terendah 0,54%, sedangkan sigmoid lebih baik pada log loss dan Brier. Ambang yang dipilih pada threshold-selection menghasilkan coverage outer-test 75,87%, accepted accuracy 95,40%, dan accepted Macro F1 95,58%. Akurasi selektif berlaku hanya pada sampel yang diterima, bukan seluruh data atau jaminan kinerja baru; kontribusi penelitian adalah evaluasi trade-off coverage–risiko, bukan peningkatan akurasi intrinsik.

Kata Kunci: Daun Singkong; CCMT; Hierarchical Classification; MobileNetV3; Random Forest; Probability Calibration; Selective Prediction.

Abstract—Field-image classification of cassava leaf disease requires trustworthy probabilities and a mechanism for withholding uncertain decisions, rather than label accuracy alone. This study evaluates HSV, GLCM, and MobileNetV3-Small embedding fusion in a two-stage hierarchical Random Forest with probability calibration and selective prediction. A total of 7,508 CCMT images were evaluated using outer stratified 5-fold cross-validation. Model comparisons report mean $\pm$ standard deviation across five outer folds. End-to-end MobileNetV3-Small achieved $93.90 \pm 0.73\%$ accuracy, $94.64 \pm 0.59\%$ balanced accuracy, and $94.23 \pm 0.73\%$ Macro F1, exceeding the Hierarchical Hybrid Random Forest at $88.28 \pm 1.12\%$, $89.66 \pm 0.76\%$, and $88.79 \pm 1.07\%$, respectively. The 1.03-percentage-point Macro F1 difference between the hierarchical and Flat Hybrid models was not significant at the 5% level (two-sided Wilcoxon, $p=0.125$). A separate calibration and selective-prediction experiment divided each outer-train fold into 70% model-fit, 15% calibrator-fit, and 15% threshold-selection partitions, all disjoint from the outer-test fold. Simplex-temperature yielded the lowest ECE (0.54%), whereas sigmoid performed better on log loss and Brier score. Thresholds selected on the held-out threshold-selection partition yielded 75.87% outer-test coverage, 95.40% accepted accuracy, and 95.58% accepted Macro F1. Selective accuracy applies only to accepted samples, not all data or guaranteed future performance; the contribution is evaluation of the coverage–risk trade-off, not improved intrinsic accuracy.

Keywords: Cassava Leaf; CCMT; Hierarchical Classification; MobileNetV3; Random Forest; Probability Calibration; Selective Prediction

1. PENDAHULUAN

Singkong (*Manihot esculenta*) merupakan salah satu tanaman pangan yang memiliki peranan penting dalam pertanian, tetapi produktivitasnya dapat terganggu oleh berbagai penyakit dan gangguan seperti bacterial blight, brown spot, green mite, dan mosaic. Identifikasi penyakit melalui pengamatan visual masih sangat dipengaruhi oleh pengalaman petani maupun penyuluh. Selain itu, gejala visual dapat memiliki kemiripan antarkelas dan dipengaruhi oleh tingkat perkembangan penyakit, kondisi pencahayaan, sudut pengambilan gambar, serta latar belakang citra. Kondisi tersebut menyebabkan klasifikasi berbasis citra menjadi permasalahan yang tidak hanya berkaitan dengan penentuan label, tetapi juga berkaitan dengan kemampuan sistem memberikan probabilitas yang dapat dipercaya ketika bukti visual yang tersedia tidak cukup kuat.

Perkembangan computer vision dan deep learning telah meningkatkan kemampuan dalam identifikasi penyakit tanaman. Dataset iCassava, misalnya, menjadi salah satu sumber data penting untuk pengembangan metode klasifikasi penyakit dan kondisi daun singkong [1]. Penelitian berbasis dataset CCMT juga menunjukkan bahwa pendekatan pembelajaran mesin dan deep learning dapat digunakan untuk mendeteksi berbagai penyakit tanaman. Mensah dkk. memperkenalkan CCMT sebagai dataset untuk crop pest and disease detection yang

mencakup beberapa tanaman dan kondisi penyakit [2]. Pada dataset tersebut, Patel dkk. menggunakan optimized recurrent neural network untuk diagnosis penyakit dan hama tanaman [3], sedangkan Jia dkk. mengembangkan pendekatan berbasis contrastive learning dan concept prototype tree untuk menghasilkan identifikasi yang lebih interpretable [4]. Pada klasifikasi khusus penyakit daun singkong, Anggraeni dkk. menggunakan transfer learning VGG16 dan memperoleh akurasi 96% [5].

Penelitian lain menunjukkan bahwa fitur warna dan tekstur tetap memiliki relevansi dalam klasifikasi penyakit tanaman. Defanka dkk. menggunakan fitur berbasis LBP dan FCH dengan Random Forest untuk klasifikasi penyakit daun jagung [6]. Huda dkk. menerapkan Gray Level Co-occurrence Matrix atau GLCM untuk klasifikasi penyakit daun padi [7]. Wildah dkk. menggabungkan Haralick, color histogram, dan Random Forest untuk klasifikasi penyakit daun kopi [8]. Pendekatan tersebut menunjukkan bahwa informasi visual sederhana seperti distribusi warna dan tekstur dapat menjadi representasi yang berguna, terutama ketika karakteristik gejala penyakit berkaitan dengan perubahan warna, bercak, atau pola permukaan daun.

Di sisi lain, perkembangan deep learning memungkinkan representasi visual dipelajari secara otomatis. Zhong dkk. menggunakan arsitektur ResNet yang dipadukan dengan transformer embedding untuk klasifikasi penyakit daun singkong pada dataset yang tidak seimbang [9]. Tran-Anh dkk. mengembangkan pendekatan multiple-model deep learning yang melibatkan MobileNetV3 dalam ensemble [10]. MobileNetV3 dirancang dengan mempertimbangkan efisiensi komputasi sehingga relevan untuk aplikasi computer vision dengan keterbatasan sumber daya [11]. Pada domain singkong, Dosset dkk. memanfaatkan MobileNetV3-Small sebagai backbone ringan yang dipadukan dengan modified soft attention untuk deteksi penyakit daun singkong [12], sedangkan Thai dkk. mengusulkan FormerLeaf sebagai vision transformer efisien untuk permasalahan yang sama [13].

HSV merepresentasikan perubahan warna daun, GLCM mengukur pola tekstur, dan embedding MobileNetV3-Small menangkap representasi visual tingkat tinggi. Fusi ketiganya diuji karena gejala dapat muncul sebagai perubahan warna, bercak permukaan, maupun pola bentuk. Informasi tersebut berpotensi saling melengkapi, tetapi manfaat fusi tidak diasumsikan: kontribusinya diuji melalui ablasi dan perbandingan handcrafted–hybrid pada protokol yang sama.

Kebutuhan tersebut kemudian tidak hanya berkaitan dengan representasi fitur, tetapi juga dengan bagaimana keputusan klasifikasi dibentuk dan seberapa dapat dipercaya probabilitas yang dihasilkan. Struktur hierarchical digunakan untuk memisahkan keputusan kondisi Healthy/Diseased dari identifikasi jenis penyakit, sehingga proses klasifikasi dilakukan secara bertahap. Namun, prediksi bertingkat tetap dapat menghasilkan confidence yang tinggi pada keputusan yang salah atau pada citra dengan bukti visual yang terbatas. Karena itu, probability calibration digunakan untuk mengevaluasi dan memperbaiki kualitas probabilitas, sedangkan selective prediction digunakan untuk menahan keputusan ketika tingkat ketidakpastian masih tinggi.

Meskipun performa klasifikasi semakin tinggi, akurasi atau F1-score belum sepenuhnya menggambarkan kualitas sistem diagnosis. Model dapat menghasilkan confidence tinggi pada prediksi yang sebenarnya salah. Dengan demikian, sistem diagnosis yang digunakan untuk mendukung pengambilan keputusan perlu memperhatikan kualitas probabilitas yang dihasilkan. Guo dkk. menunjukkan bahwa neural network modern dapat menghasilkan probabilitas yang tidak terkalibrasi dengan baik dan bahwa post-hoc calibration dapat digunakan untuk memperbaiki hubungan antara confidence dan frekuensi kebenaran prediksi [14]. Konsep probability calibration juga telah dikaji dalam machine learning klasik, termasuk pendekatan supervised probability prediction [15].

Permasalahan lain muncul ketika sistem tidak memiliki bukti yang cukup untuk memberikan keputusan pada seluruh sampel. Dalam kondisi tersebut, sistem dapat diberikan kemampuan untuk melakukan selective prediction, yaitu menerima prediksi yang dianggap cukup meyakinkan dan menolak atau menahan prediksi yang memiliki ketidakpastian tinggi. Geifman dan El-Yaniv memperkenalkan selective classification untuk deep neural networks sebagai pendekatan yang mengevaluasi hubungan antara risiko kesalahan dan coverage, yaitu proporsi sampel yang tetap diberikan keputusan [16]. Dengan pendekatan tersebut, sistem tidak dipaksa memberikan diagnosis pada seluruh sampel.

Struktur hierarkis juga dapat digunakan untuk memberikan keputusan secara bertahap. Goren dkk. mengembangkan konsep hierarchical selective classification yang mempertimbangkan struktur hierarki dalam proses selective decision [17]. Konsep tersebut relevan dengan diagnosis penyakit daun singkong karena permasalahan klasifikasi dapat dipisahkan menjadi dua tingkat. Tingkat pertama menentukan apakah daun berada dalam kondisi sehat atau sakit. Apabila daun dikategorikan sakit, tingkat kedua menentukan jenis penyakit. Struktur tersebut memungkinkan sistem mempertahankan informasi bahwa sebuah daun terindikasi sakit tanpa harus selalu memaksakan diagnosis penyakit tertentu ketika bukti pada tahap kedua tidak cukup kuat.

Berdasarkan penelitian-penelitian tersebut, masih terdapat kebutuhan untuk mengevaluasi representasi fitur, struktur keputusan, dan kualitas probabilitas secara terintegrasi pada klasifikasi penyakit daun singkong. Penelitian sebelumnya telah mengeksplorasi pendekatan deep learning, fitur warna dan tekstur, serta metode klasifikasi penyakit singkong, tetapi evaluasi yang secara bersamaan memisahkan pengaruh fusi representasi handcrafted dan deep feature, struktur flat dan hierarchical, kalibrasi probabilitas, serta kemampuan selective prediction masih terbatas. Celah ini penting karena akurasi label yang tinggi tidak otomatis menunjukkan bahwa probabilitas model terkalibrasi atau bahwa keputusan dengan ketidakpastian tinggi dapat ditahan secara andal.

Untuk menjawab celah tersebut, penelitian ini mengevaluasi Hierarchical Hybrid Random Forest, yaitu Random Forest dengan struktur keputusan hierarkis Healthy/Diseased menuju kelas penyakit yang menggunakan fusi fitur handcrafted HSV-GLCM dan deep feature embedding MobileNetV3-Small. Struktur flat dan hierarchical dibandingkan pada outer fold yang sama. MobileNetV3-Small end-to-end digunakan sebagai baseline prediktif untuk mengontekstualisasikan performa klasifikasi, bukan untuk menyatakan keunggulan Hierarchical Hybrid Random Forest terhadap deep learning end-to-end. Probabilitas Random Forest dievaluasi tanpa kalibrasi serta dengan sigmoid, isotonic, dan simplex-temperature. Prediksi selektif memilih ambang hanya pada partisi threshold-selection sebelum diterapkan pada outer-test independen.

Kontribusi penelitian terletak pada evaluasi terintegrasi yang memisahkan tiga pertanyaan: pengaruh representasi fitur, pengaruh struktur keputusan flat versus hierarchical, dan hubungan antara kalibrasi probabilitas dengan kemampuan abstention. Kontribusi ini diposisikan sebagai rancangan evaluasi empiris, bukan sebagai klaim bahwa komponen HSV, GLCM, MobileNetV3, Random Forest, atau kalibrasi merupakan algoritma baru.

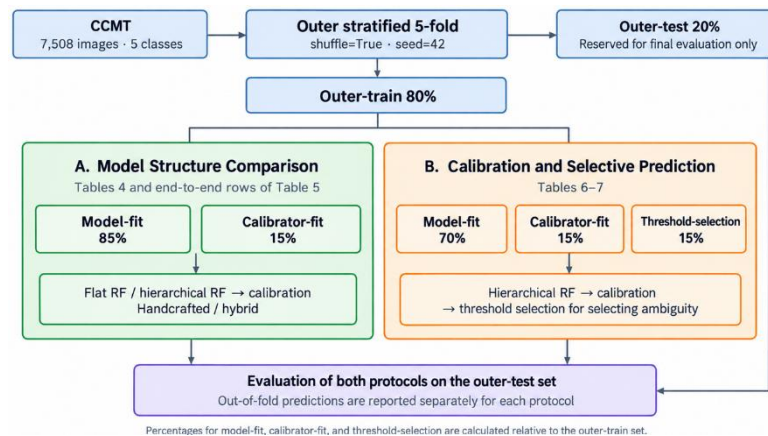
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian meliputi akuisisi dan pra-pemrosesan citra, ekstraksi HSV dan GLCM, ekstraksi embedding MobileNetV3-Small, fusi fitur, pelatihan Random Forest flat dan hierarkis, pelatihan baseline MobileNetV3-Small end-to-end, kalibrasi probabilitas, pemilihan ambang prediksi selektif, serta evaluasi. Urutan ini memisahkan pengujian representasi fitur, struktur keputusan, kualitas probabilitas, dan trade-off coverage–risiko.

Struktur hierarchical menggunakan dua classifier. Tahap pertama merupakan binary classifier yang membedakan Healthy dan Diseased. Citra yang diprediksi sebagai Diseased diteruskan ke tahap kedua untuk menentukan empat kelas penyakit, yaitu bacterial blight, brown spot, green mite, dan mosaic. Struktur tersebut memungkinkan keputusan dilakukan secara kondisional.

Gambar 1 membedakan dua pembagian outer-train dalam outer stratified 5-fold: 85% model-fit dan 15% calibrator-fit untuk perbandingan struktur RF (Tabel 4; end-to-end Tabel 5), serta 70% model-fit, 15% calibrator-fit, dan 15% threshold-selection untuk kalibrasi dan prediksi selektif (Tabel 6–7). Outer-test hanya untuk evaluasi, bukan fitting atau pemilihan konfigurasi. Stage 1–2 Tabel 5 memakai holdout 70%/15%/15% pada dataset yang sama, bukan validasi eksternal.



Gambar 1. Dua protokol outer 5-fold untuk perbandingan struktur RF serta kalibrasi dan prediksi selektif.

2.2 Dataset dan Pra-pemrosesan

Dataset yang digunakan merupakan subset singkong dari CCMT: Dataset for Crop Pest and Disease Detection, versi 1, dengan DOI 10.17632/bwh3zbpkp.v1 [2]. Citra dikumpulkan menggunakan kamera resolusi tinggi pada lahan lokal di Ghana dengan variasi dimensi, pencahayaan, sudut pengambilan, dan latar belakang. Label citra telah divalidasi oleh ahli virologi tanaman. Penelitian ini menggunakan 7.508 citra asli dan tidak menggunakan 26.330 citra hasil augmentasi sebagai data awal. Penggunaan citra asli bertujuan menghindari kemungkinan transformasi dari citra sumber yang sama tersebar pada subset data yang berbeda. Dataset terdiri atas lima kelas, yaitu bacterial blight, brown spot, green mite, healthy, dan mosaic, dengan distribusi seperti ditunjukkan pada Tabel 1.

Tabel 1. Distribusi 7.508 Citra Mentah Subset Singkong CCMT

Kelas	Jumlah Citra	Persentase
Bacterial Blight	2.614	34,82%
Brown Spot	1.481	19,73%

Kelas	Jumlah Citra	Persentase
Green Mite	1.015	13,52%
Healthy	1.193	15,89%
Mosaic	1.205	16,05%
Total	7.508	100%

Rasio jumlah kelas terbesar terhadap kelas terkecil adalah sekitar 2,58:1 sehingga ketidakseimbangan kelas dikategorikan moderat. Seluruh citra diubah menjadi ukuran 224×224 piksel sebelum proses ekstraksi fitur dan pelatihan model.

Pembagian menggunakan stratified split pada tingkat citra untuk mempertahankan proporsi kelas. Dataset tidak menyediakan identitas daun, tanaman, sesi pengambilan, atau lokasi, sehingga group-aware splitting tidak dapat dilakukan. Citra yang berkorelasi dari daun atau lokasi yang sama mungkin tersebar antarfold dan meningkatkan estimasi performa. Oleh sebab itu, hasil penelitian dibatasi sebagai validasi internal tingkat citra dan tidak dianggap sebagai bukti generalisasi ke tanaman, lokasi, atau musim baru. Validasi lapangan independen atau dataset dengan group identifier diperlukan untuk mengatasi risiko leakage ini.

2.3 Ekstraksi Fitur Warna

Fitur warna diekstraksi menggunakan ruang warna HSV (Hue, Saturation, Value) yang dipilih untuk merepresentasikan karakteristik warna citra pada kondisi pencahayaan lapangan yang beragam. Dua kelompok deskriptor digunakan, yaitu color moments dan histogram warna [18].

Color moments dihitung pada masing-masing kanal HSV menggunakan tiga statistik, yaitu mean, standar deviasi (standard deviation), dan skewness. Perhitungan ketiga statistik tersebut ditunjukkan pada Persamaan (1)–(3), dengan (x_i) menyatakan nilai piksel pada kanal tertentu dan (N) merupakan jumlah piksel.

$$\mu = (1/N) \sum x_i \quad (1)$$

$$\sigma = \sqrt{[(1/N) \sum (x_i - \mu)^2]} \quad (2)$$

$$s = (1/N) \sum [(x_i - \mu)/\sigma]^3 \quad (3)$$

Selain color moments, digunakan histogram warna tiga dimensi pada ruang HSV dengan konfigurasi $8 \times 8 \times 8$ bin. Histogram dinormalisasi sebelum digunakan sebagai fitur. Dengan demikian, representasi warna terdiri atas 9 fitur color moments dan 512 fitur histogram, sehingga menghasilkan total 521 fitur warna.

2.4 Ekstraksi Fitur Tekstur

Fitur tekstur diekstraksi menggunakan Gray Level Co-occurrence Matrix (GLCM), yang merepresentasikan frekuensi kemunculan pasangan intensitas piksel yang berdekatan pada jarak dan orientasi tertentu [7], [19]. Sebelum pembentukan GLCM, citra dikonversi ke grayscale dan dikuantisasi menjadi 32 tingkat intensitas untuk mengurangi kompleksitas matriks.

GLCM dihitung pada jarak satu piksel dengan empat orientasi, yaitu 0° , 45° , 90° , dan 135° . Nilai dari keempat orientasi tersebut kemudian dirata-ratakan untuk mengurangi sensitivitas fitur terhadap orientasi, meskipun proses tersebut tidak menjamin invariansi rotasi secara penuh. Enam properti tekstur digunakan, yaitu contrast, dissimilarity, homogeneity, energy, correlation, dan Angular Second Moment (ASM). Salah satu perhitungan ditunjukkan pada Persamaan (4), dengan $(P(i,j))$ merupakan elemen GLCM yang telah dinormalisasi.

$$\text{Contrast} = \sum_i \sum_j (i - j)^2 P(i,j) \quad (4)$$

Fitur tekstur digunakan untuk melengkapi informasi warna karena gejala penyakit seperti bercak, mosaik, dan perubahan permukaan daun dapat menghasilkan perbedaan pola tekstur [6], [8]. Hasil ekstraksi menghasilkan enam fitur GLCM.

2.5 Fusi Handcrafted Feature dan Deep Feature

Fitur HSV sebanyak 521 dimensi merepresentasikan distribusi warna, sedangkan enam fitur GLCM mengukur tekstur spasial. Keduanya membentuk 527 fitur handcrafted. Embedding MobileNetV3-Small prelatih ImageNet menambahkan 576 dimensi yang menangkap bentuk dan konteks visual. Concatenation menghasilkan 1.103 dimensi. Komplementaritas ketiga kelompok tersebut diperlakukan sebagai hipotesis dan diuji melalui ablasi, bukan diasumsikan sebagai peningkatan otomatis.

Vektor fitur tersebut digunakan sebagai masukan Random Forest tanpa StandardScaler. Hal ini menyederhanakan pipeline karena Random Forest tidak memerlukan standarisasi fitur seperti model yang sensitif terhadap skala atau berbasis optimasi gradien. Pemisahan data dilakukan sebelum proses balancing dan pelatihan untuk menjaga independensi data evaluasi.

Eksperimen menggunakan outer stratified 5-fold cross-validation pada tingkat citra dengan shuffle dan seed 42. Perbandingan struktur model pada Tabel 4 memakai 85% model-fit dan 15% calibrator-fit di dalam outer-train karena eksperimen tersebut hanya membandingkan struktur flat dan hierarkis tanpa memilih ambang; pembagian



ini mempertahankan lebih banyak sampel pelatihan dan konsisten dengan protokol perbandingan sebelumnya. Tabel 6–7 menjawab tujuan yang berbeda, yaitu evaluasi kalibrasi dan prediksi selektif yang memerlukan partisi pemilihan ambang independen. Karena itu, outer-train pada Tabel 6–7 dibagi menjadi 70% model-fit, 15% calibrator-fit, dan 15% threshold-selection; persentase ini relatif terhadap outer-train, bukan seluruh dataset. Partisi terakhir untuk memilih ambang, bukan untuk fitting calibrator atau pelatihan model. Metrik kedua protokol dilaporkan terpisah dan tidak ditafsirkan sebagai perbandingan performa langsung. Seluruh pembagian dilakukan secara stratified dengan seed turunan fold. Outer-test tidak digunakan untuk fitting, kalibrasi, early stopping, pemilihan calibrator, atau pemilihan ambang. Random Forest menggunakan 500 pohon, kriteria Gini, $\text{max_features}=\text{sqrt}$, $\text{class_weight}=\text{balanced_subsample}$, dan seed turunan fold. Risiko korelasi antar citra tetap ada karena group identifier tidak tersedia.

MobileNetV3-Small end-to-end digunakan sebagai baseline prediktif untuk mengontekstualisasikan performa klasifikasi, bukan sebagai baseline untuk seluruh pipeline kalibrasi-prediksi selektif dan bukan untuk mendukung klaim keunggulan Random Forest atas deep learning. Model diinisialisasi dengan bobot ImageNet dan seluruh layer di-fine-tune maksimal lima epoch menggunakan AdamW (learning rate 2×10^{-4} ; weight decay 1×10^{-4}), batch size 32, weighted cross-entropy, dan cosine annealing. Augmentasi pada training partition meliputi random resized crop 224×224 (scale 0,75–1,00), horizontal flip, rotasi $\pm 15^\circ$, serta color jitter (brightness 0,15; contrast 0,15; saturation 0,10). Checkpoint dipilih berdasarkan Macro F1 pada validation partition sebesar 15% dari outer-train, dengan early stopping patience dua epoch. Validation partition ini hanya digunakan untuk pemilihan checkpoint dan early stopping CNN; fungsinya berbeda dari calibrator-fit pada pipeline Random Forest dan tidak digunakan untuk memilih threshold. Outer-test hanya digunakan satu kali untuk evaluasi.

Tabel 2. Rincian Dimensi Fitur yang Diekstraksi dan Fusi

Kelompok Fitur	Deskriptor	Jumlah Dimensi
Warna (HSV)	Color moments (mean, std, skewness $\times$ 3 kanal)	9
Warna (HSV)	Histogram warna $8 \times 8 \times 8$ bin	512
Tekstur (GLCM)	Contrast, dissimilarity, homogeneity, energy, correlation, ASM	6
Deep feature	MobileNetV3-Small pretrained ImageNet embedding	576
Total Hybrid Feature	HSV + GLCM + MobileNetV3 (concatenation)	1.103

Representasi fitur yang digunakan terdiri atas sembilan color moments dan 512 histogram HSV, enam fitur GLCM, serta 576 embedding MobileNetV3-Small. Dengan demikian, total dimensi pada Hybrid Feature adalah 1.103.

Istilah konfigurasi Random Forest dibedakan secara tegas. Pada studi ablasi Tabel 3, RF Standard menggunakan 200 pohon tanpa class_weight , resampling, atau calibrator-fit; model dilatih pada seluruh outer-train. RF Balanced menggunakan 200 pohon tanpa class_weight dan menerapkan over/undersampling berbasis median hanya pada data latih. Perbandingan struktur pada Tabel 4–5 menggunakan 500 pohon, $\text{max_features}=\text{sqrt}$, $\text{class_weight}=\text{'balanced_subsample'}$, resampling pada model-fit, serta pembagian 85% model-fit dan 15% calibrator-fit. Evaluasi Kalibrasi dan Prediksi Selektif pada Tabel 6–7 menggunakan konfigurasi RF yang sama dengan pembagian 70% model-fit, 15% calibrator-fit, dan 15% threshold-selection [20].

Pada seluruh prosedur yang menggunakan resampling, kelas di bawah median di-oversample secara acak dengan replacement dan kelas di atas median di-undersample tanpa replacement. Calibrator-fit, threshold-selection, dan outer-test mempertahankan distribusi asli. Karena Tabel 3 dan Tabel 4 berbeda dalam jumlah pohon, class_weight , resampling, ukuran model-fit, dan kalibrasi, Fusion RF Standard 89,12% pada Tabel 3 tidak dibandingkan langsung dengan Flat Handcrafted 86,47% pada Tabel 4. Tabel 3 menguji kelompok fitur dan resampling pada baseline, sedangkan Tabel 4 membandingkan struktur keputusan dalam protokol RF utama.

2.6 Kalibrasi, Pemilihan Operating Point, dan Evaluasi

Random Forest menghasilkan probabilitas dari proporsi vote pohon. Kalibrasi dipasang secara terpisah pada classifier kesehatan dan classifier penyakit menggunakan calibrator-fit. Untuk simplex-temperature, probabilitas mentah dipotong ke rentang $[10^{-12}, 1]$, lalu ditransformasikan dengan $p'_i = p_i^{(1/T)} / \sum_j p_j^{(1/T)}$. Nilai T dicari secara independen untuk setiap classifier dan fold dengan `scipy.optimize.minimize_scalar` metode bounded pada rentang 0,25–5,00, menggunakan mean negative log-likelihood pada calibrator-fit sebagai fungsi objektif. Sigmoid dan isotonic diterapkan melalui CalibratedClassifierCV pada estimator prefit. Pada classifier penyakit empat kelas, calibrator dipasang secara one-vs-rest untuk setiap kelas, kemudian probabilitas dinormalisasi agar berjumlah satu. Probabilitas akhir lima kelas dibentuk sebagai $P(\text{Healthy})$ untuk kelas sehat dan $P(\text{Diseased}) \times P(k|\text{Diseased})$ untuk setiap penyakit, lalu dinormalisasi kembali. ECE dihitung dengan 10 equal-width bins, sedangkan multiclass Brier score merupakan rata-rata jumlah kuadrat selisih probabilitas lima kelas terhadap target one-hot.

Simplex-temperature digunakan sebagai konfigurasi operasional tetap dalam kode untuk perhitungan operating point, bukan sebagai hasil seleksi otomatis berdasarkan skor outer-test. Pada Tabel 6–7, temperatur dipelajari hanya pada calibrator-fit dan ambang dipilih hanya dari probabilitas simplex-temperature pada threshold-selection. Sigmoid dan isotonic dievaluasi sebagai pembanding, tanpa aturan yang mengganti calibrator

operasional berdasarkan ECE, log loss, atau Brier outer-test. Artefak eksperimen tidak memuat protokol prapenetapan bertanggal; karena itu, penggunaan metode tetap dalam kode tidak dianggap sebagai bukti bahwa pilihan awal dibuat sebelum seluruh hasil uji pernah ditinjau.

Untuk evaluasi pada Subbab 3.7 dan Tabel 7, confidence adalah probabilitas terbesar pada vektor akhir lima kelas dan margin adalah selisih dua probabilitas terbesar. Sampel diterima hanya jika confidence $\geq$ tau dan margin $\geq$ delta. Grid tau=0,50–0,95 dengan langkah 0,01 dan delta=0,00–0,40 dengan langkah 0,02 dievaluasi hanya pada threshold-selection. Kandidat yang mencapai accepted accuracy minimal 95% diurutkan berdasarkan coverage tertinggi, accepted accuracy tertinggi, tau terendah, dan delta terendah. Seluruh fold mencapai target pada threshold-selection. Pasangan terpilih kemudian diterapkan satu kali pada outer-test; keputusan parsial 'tidak sehat tanpa jenis penyakit' tetap dihitung abstained. Jika delta=0,00, syarat margin otomatis terpenuhi karena selisih dua probabilitas terbesar tidak negatif; penerimaan efektif ditentukan oleh confidence.

Calibrator dan ambang selective prediction tidak memakai partisi yang sama. Audit indeks pada lima fold menunjukkan overlap nol antara model-fit, calibrator-fit, threshold-selection, dan outer-test. Pemisahan ini mencegah reuse sampel antara calibration dan threshold selection serta menjaga outer-test tetap independen dari penetapan pipeline. Namun, pemisahan pada tingkat citra tidak menghilangkan kemungkinan group-level leakage karena identitas daun, tanaman, sesi, dan lokasi tidak tersedia.

$$\text{Akurasi} = \text{jumlah prediksi benar} / \text{jumlah seluruh sampel} \quad (5)$$

$$\text{Presisi} = \text{TP} / (\text{TP} + \text{FP}) \quad (6)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (7)$$

$$\text{F1-score} = 2 \times (\text{Presisi} \times \text{Recall}) / (\text{Presisi} + \text{Recall}) \quad (8)$$

Presisi, recall, dan F1 pada Persamaan (6)-(8) dihitung untuk setiap kelas dengan skema one-vs-rest. Macro F1 adalah rata-rata aritmetika F1 kelima kelas, sehingga setiap kelas memiliki bobot yang sama tanpa dipengaruhi jumlah sampelnya. Balanced accuracy adalah rata-rata recall kelima kelas.

Analisis SHAP diterapkan hanya pada baseline Fusion RF handcrafted. Nama hist_hX_sY_vZ menyatakan bin histogram HSV; misalnya hist_h0_s3_v3 merepresentasikan hue 0-22,5, saturation 96-128, dan value 96-128. Interpretasi menghubungkan fitur dengan pola perubahan warna yang mungkin terkait klorosis, mosaik, atau lesi, tetapi SHAP hanya menunjukkan kontribusi prediktif. Analisis ini tidak merepresentasikan feature attribution pada Hierarchical Hybrid karena 576 embedding MobileNetV3 diperlakukan sebagai latent representation dan tidak dianalisis secara semantik [21].

2.7 Implementasi Sistem Hierarkis

Pipeline klasifikasi hierarkis diimplementasikan sebagai sistem inferensi untuk mendemonstrasikan penerapan model. Dua artefak Random Forest yang telah dikalibrasi disimpan dalam format joblib. Backend berbasis FastAPI melakukan pra-pemrosesan citra dan ekstraksi 1.103 fitur tanpa StandardScaler, kemudian menghasilkan probabilitas tahap Healthy–Diseased dan probabilitas kondisional untuk empat kelas penyakit.

Sistem inferensi menerapkan dua gerbang keputusan. Gerbang Tahap 1 menggunakan confidence dan margin pada probabilitas Healthy–Diseased untuk menerima atau menahan status kesehatan. Jika status Diseased diterima, gerbang diagnosis menggunakan confidence dan margin dari probabilitas akhir lima kelas. Dengan demikian, sistem dapat mempertahankan keluaran 'tidak sehat' ketika Tahap 1 diterima, tetapi menahan jenis penyakit ketika gerbang diagnosis gagal. Aturan parsial ini digunakan untuk komunikasi antarmuka, sedangkan accepted pada evaluasi utama tetap berarti label akhir lima kelas diterima.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Studi Ablasi Kelompok Fitur

Enam skema baseline dievaluasi melalui stratified 5-fold cross-validation pada 7.508 citra. Akurasi, balanced accuracy, Macro F1, dan weighted F1 pada Tabel 3 dihitung dari gabungan seluruh prediksi out-of-fold, bukan rerata metrik antar-fold. Kolom SD Macro F1 dihitung dari lima skor Macro F1 fold menggunakan simpangan baku sampel (ddof=1; penyebut n–1=4), dan disajikan terpisah dari Macro F1 agregat, bukan sebagai rerata $\pm$ SD. Karena hanya tersedia lima fold dan citra mungkin berkorelasi, variasi fold serta uji statistik diinterpretasikan secara eksploratif, bukan bukti definitif superioritas konfigurasi.

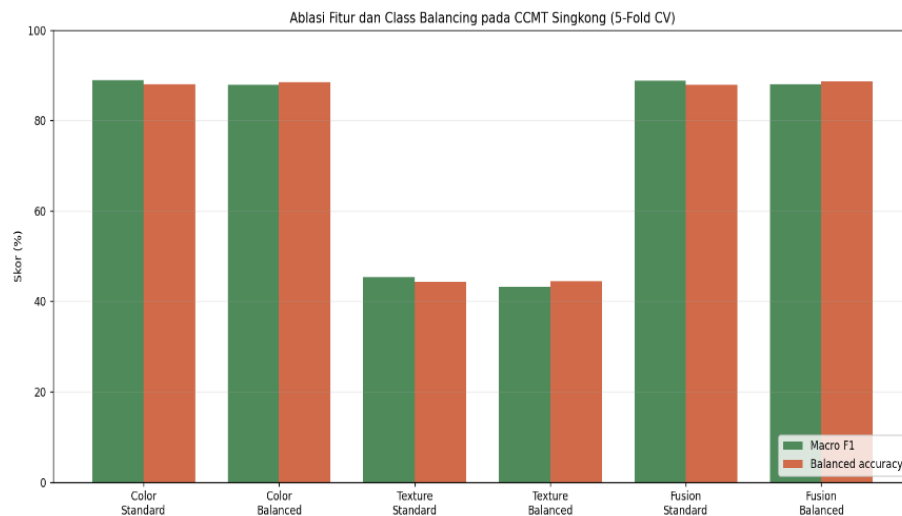
Tabel 3. Perbandingan Kelompok Fitur dan Skema Random Forest (Prediksi OOF, 5-Fold Cross-Validation)

Kelompok Fitur	Model	Akurasi	Bal. Acc.	Macro F1	SD Macro F1	Weighted F1
Color-only	RF Standard	89,09%	88,03%	88,90%	0,96%	89,07%
Color-only	RF Balanced	88,15%	88,45%	87,96%	1,22%	88,20%
Texture-only	RF Standard	48,15%	44,23%	45,30%	0,49%	47,35%

Kelompok Fitur	Model	Akurasi	Bal. Acc.	Macro F1	SD Macro F1	Weighted F1
Texture-only	RF Balanced	43,54%	44,42%	43,27%	0,87%	43,64%
Fusion	RF Standard	89,12%	87,92%	88,86%	1,01%	89,10%
Fusion	RF Balanced	88,24%	88,64%	88,10%	0,90%	88,29%

Color-only RF Standard memperoleh akurasi 89,09%, balanced accuracy 88,03%, dan Macro F1 88,90%. Fusion RF Standard menghasilkan akurasi tertinggi baseline sebesar 89,12%, sedangkan Fusion RF Balanced mencapai balanced accuracy 88,64% dan Macro F1 88,10%. Texture-only tertinggal dengan Macro F1 45,30% (Standard) dan 43,27% (Balanced). Dengan demikian, warna merupakan sinyal handcrafted dominan pada CCMT, sedangkan penambahan enam fitur GLCM hanya memberi perubahan kecil terhadap model warna.

Gambar 2 menunjukkan Color-only dan Fusion berada pada kisaran kinerja yang berdekatan, sedangkan Texture-only tertinggal jauh. RF Balanced sedikit meningkatkan balanced accuracy pada Color-only, Texture-only, dan Fusion, tetapi tidak selalu meningkatkan Macro F1 atau akurasi. Visual ini menegaskan bahwa efek balancing pada CCMT bersifat trade-off, bukan peningkatan universal.



Gambar 2. Perbandingan Macro F1 dan Balanced Accuracy pada Studi Ablasi Fitur dan Resampling (5-Fold CV)

3.2 Evaluasi Pengaruh Resampling pada Ketidakseimbangan Moderat

Distribusi CCMT memiliki rasio kelas maksimum 2,58:1, sehingga balancing tidak otomatis diperlukan. Fusion RF Balanced menurunkan accuracy dari 89,12% menjadi 88,24%, tetapi menaikkan balanced accuracy dari 87,92% menjadi 88,64%; Macro F1 juga turun dari 88,86% menjadi 88,10%. Uji Friedman pada enam konfigurasi menunjukkan perbedaan keseluruhan ($\chi^2=23,63$; $p<0,001$; Kendall's $W=0,945$), tetapi perbandingan terpilih dengan Wilcoxon dan koreksi Holm belum signifikan pada $\alpha 0,05$. Oleh karena itu, resampling dipertahankan sebagai ablasi dan keputusan konfigurasi harus didasarkan pada trade-off accuracy, Macro F1, balanced accuracy, serta recall per kelas.

3.3 Recall per Kelas

Recall per kelas pada konfigurasi fusion menunjukkan trade-off yang sama dengan metrik agregat. Fusion RF Standard menghasilkan recall 94,72% untuk bacterial blight, 84,67% untuk brown spot, 86,31% untuk green mite, 88,43% untuk healthy, dan 85,48% untuk mosaic. Fusion RF Balanced menurunkan recall bacterial blight menjadi 85,96%, tetapi meningkatkan brown spot menjadi 90,61%, green mite 87,49%, healthy 89,10%, dan mosaic 90,04%. Hasil ini menunjukkan bahwa resampling memperbaiki beberapa kelas tanpa meningkatkan seluruh kelas secara seragam.

3.4 Confusion Matrix: Fusion RF Standard vs. RF Balanced

Untuk menjaga fokus pembahasan, dua confusion matrix penuh tidak ditampilkan sebagai tabel terpisah. Pola utamanya menunjukkan bahwa RF Standard mempertahankan recall bacterial blight yang lebih tinggi, sedangkan resampling meningkatkan pengenalan brown spot dan mosaic. Temuan ini konsisten dengan trade-off balanced accuracy dan accuracy pada Tabel 3 dan ringkasan recall pada Subbab 3.3.

Secara agregat, Fusion Standard mengklasifikasikan benar 2.476 dari 2.614 bacterial blight. Fusion Balanced menurunkannya menjadi 2.247, tetapi meningkatkan prediksi benar brown spot dari 1.254 menjadi 1.342 dan mosaic dari 1.030 menjadi 1.085. Ringkasan ini menjelaskan mengapa resampling menaikkan balanced accuracy tetapi sedikit menurunkan accuracy total.

3.5 Analisis Interpretabilitas Model Menggunakan SHAP

Analisis SHAP diterapkan pada baseline Fusion RF handcrafted untuk memeriksa sumber atribusi prediktif. Secara agregat, fitur warna menyumbang 90,63% dari jumlah mean absolute SHAP dan enam GLCM 9,37% pada model tersebut. Nilai ini menjelaskan baseline handcrafted, bukan kontribusi kausal dan bukan feature attribution pada Hierarchical Hybrid final.

Fitur dengan atribusi terbesar terutama berasal dari histogram saturation–value dan S_mean. Dominasi ini mendukung temuan ablasi bahwa warna merupakan sinyal handcrafted utama pada CCMT, tetapi dapat pula dipengaruhi pencahayaan dan latar belakang. Analisis embedding dan Grad-CAM diperlukan untuk menjelaskan komponen deep feature.

3.6 Evaluasi Flat versus Hierarchical pada Tingkat Sistem

Tabel 4 membandingkan model pada outer fold identik (Subbab 2.5). Hierarchical Hybrid memperoleh rerata Macro F1 lebih tinggi 1,03 poin persentase daripada Flat Hybrid. Uji Wilcoxon signed-rank exact dua sisi terhadap lima pasangan skor Macro F1 menghasilkan $W=1$ dan $p=0,125$. Karena $p>0,05$, perbedaannya belum signifikan secara statistik pada taraf 5%. Hasil ini menunjukkan keunggulan deskriptif model hierarkis, tetapi belum memberikan bukti statistik yang cukup untuk menyimpulkan bahwa model hierarkis lebih baik daripada model datar.

Tabel 4. Perbandingan Kinerja Model pada Outer 5-Fold (Rerata $\pm$ SD).

Model	Accuracy	Bal Accuracy	Macro F1	Keterangan
Flat Handcrafted	86,47 $\pm$ 1,08%	86,90 $\pm$ 0,81%	86,24 $\pm$ 1,06%	527 fitur
Hierarchical Handcrafted	87,39 $\pm$ 1,44%	87,56 $\pm$ 1,36%	87,22 $\pm$ 1,41%	527 fitur
Flat Hybrid	87,24 $\pm$ 1,74%	89,03 $\pm$ 1,44%	87,76 $\pm$ 1,71%	1.103 fitur
Hierarchical Hybrid	88,28 $\pm$ 1,12%	89,66 $\pm$ 0,76%	88,79 $\pm$ 1,07%	1.103 fitur
MobileNetV3-Small end-to-end	93,90 $\pm$ 0,73%	94,64 $\pm$ 0,59%	94,23 $\pm$ 0,73%	Baseline CNN; tanpa kalibrasi RF

Catatan: RF menggunakan 500 pohon, resampling, class_weight='balanced_subsample', serta 85% model-fit dan 15% calibrator-fit. Protokol 70%/15%/15% untuk kalibrasi dan prediksi selektif dilaporkan terpisah pada Tabel 6–7

3.7 Interpretasi Metrik Tahap dan Keputusan End-to-End

Metrik tahap dan metrik sistem menjawab pertanyaan yang berbeda. Evaluasi stage-level pada 1.127 citra holdout digunakan khusus untuk mengisolasi performa setiap classifier dalam struktur hierarkis, sedangkan evaluasi utama memakai seluruh 7.508 prediksi out-of-fold dari outer 5-fold untuk mengukur performa end-to-end. Stage 1 merupakan skrining Healthy versus Diseased. Stage 2 hanya menilai 948 citra holdout berlabel Diseased, sehingga metriknya bersifat kondisional, tidak memasukkan kesalahan Stage 1, dan bukan accuracy akhir sistem. Holdout stage-level membagi 7.508 citra yang sama menjadi 5.255 model-fit, 1.126 calibrator-fit, dan 1.127 test (70%/15%/15%, seed 42). Holdout dan outer 5-fold adalah dua protokol pada dataset yang sama, bukan validasi eksternal atau bukti validasi independen; seluruh citra holdout-test juga menjadi outer-test dalam salah satu fold CV. Tabel 5 memisahkan hasil diagnostik tiap tahap dari hasil sistem utama.

Tabel 5. Perbedaan Metrik Stage 1, Stage 2 Kondisional, dan End-to-End

Evaluasi	Data evaluasi	Balanced Acc.	Macro F1	Accuracy	Interpretasi
Stage 1	1.127 holdout; Healthy vs Diseased	94,48%	96,05%	97,96%	Recall diseased 99,58%; healthy 89,39%
Stage 2 kondisional	948 holdout berlabel Diseased	90,64%	89,42%	89,35%	P(diagnosis benar Diseased)
End-to-end utama	7.508 OOF; 5-fold; lima kelas	89,66 $\pm$ 0,76%	88,79 $\pm$ 1,07%	88,28 $\pm$ 1,12%	Mencakup propagasi kesalahan dua tahap

Untuk Stage 1, balanced accuracy dan Macro F1 lebih ditonjolkan daripada accuracy karena kelas Diseased jauh lebih banyak daripada Healthy. Selisih recall menunjukkan bahwa model sengaja lebih sensitif terhadap daun sakit, tetapi konsekuensinya sebagian daun sehat dapat ditandai sakit. Karena itu, angka accuracy 97,96% selalu dilaporkan bersama recall kedua kelas dan tidak digunakan sebagai klaim performa utama.

Analisis demonstratif memakai pasangan tetap confidence 0,80 dan margin 0,20 untuk memperlihatkan distribusi keputusan end-to-end. Pasangan ini bukan operating point final dan tidak digunakan untuk menyimpulkan performa deployment. Sampel dihitung accepted hanya jika kedua syarat terpenuhi; keputusan parsial status sakit tanpa jenis penyakit tetap dihitung abstained. Operating point hasil selection per fold pada threshold-selection yang independen dilaporkan pada Tabel 7.

Pada threshold demonstratif, 1.015 citra Healthy diklasifikasikan benar dan diterima, 16 salah diterima sebagai Diseased, dan 162 ditahan. Pada kelompok Diseased, 4.352 citra menerima jenis penyakit yang benar, 221 menerima kelas akhir yang salah, dan 1.742 ditahan. Secara keseluruhan, 5.604 dari 7.508 prediksi (74,64%) diterima dan 1.904 (25,36%) ditahan; 5.367 keputusan yang diterima benar, sehingga accepted accuracy mencapai 95,77%. Nilai tersebut menggambarkan threshold tetap untuk demonstrasi, bukan operating point hasil selection.

3.8 Evaluasi Kalibrasi Probabilitas

Tabel 6 membandingkan kalibrasi Hierarchical Hybrid pada protokol tiga partisi. Model hanya dilatih pada model-fit, calibrator hanya dipasang pada calibrator-fit, dan evaluasi memakai outer-test terpisah. Metrik dihitung setelah menggabungkan 7.508 prediksi out-of-fold dari lima outer fold; setiap citra menyumbang satu prediksi. Nilainya adalah agregat pooled OOF, bukan rerata $\pm$ standar deviasi antar-fold. ECE memakai 10 equal-width bins pada confidence gabungan. Macro F1 88,79 $\pm$ 1,07% pada Tabel 4 dan 88,17% pada Tabel 6 tidak dibandingkan sebagai peningkatan atau penurunan model: Tabel 4 memakai 85% model-fit untuk perbandingan struktur, sedangkan Tabel 6–7 memakai 70% model-fit karena menyisihkan threshold-selection. Ukuran pelatihan dan tujuan evaluasi berbeda.

Tabel 6. Perbandingan Metode Kalibrasi Hierarchical Hybrid dengan Protokol Tiga Partisi (Pooled OOF, n=7.508; Acc.=Accuracy; Bal. Acc.=Balanced Accuracy)

Kondisi	Acc.	Log Loss	Brier	ECE	Bal. Acc.	Macro F1
Uncalibrated	87,28%	0,5642	0,2743	23,57%	88,88%	87,73%
Simplex-temp.	87,61%	0,3414	0,1803	0,54%	89,17%	88,17%
Sigmoid	88,44%	0,3286	0,1698	2,87%	88,91%	88,77%
Isotonic	88,36%	0,454	0,1735	1,37%	88,88%	88,74%

Simplex-temperature mengubah accuracy dari 87,28% menjadi 87,61% serta menurunkan ECE dari 23,57% menjadi 0,54%. Sigmoid menghasilkan accuracy 88,44%, log loss 0,3286, Brier 0,1698, dan Macro F1 88,77%, sedangkan isotonic menghasilkan ECE 1,37% dan Macro F1 88,74%. Perbandingan menunjukkan trade-off antarmetrik. Simplex-temperature merupakan konfigurasi operasional tetap yang diaudit pada kode, bukan metode yang dinyatakan terbaik berdasarkan seluruh metrik atau hasil seleksi outer-test.

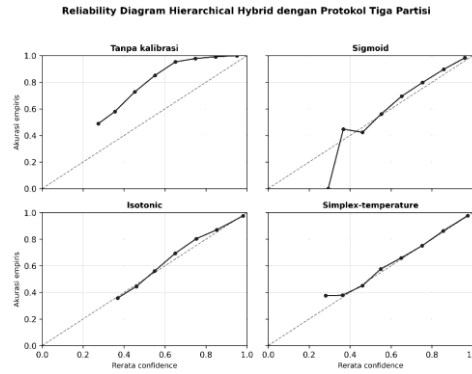
3.9 Evaluasi Prediksi Selektif

Berbeda dari analisis demonstratif pada Subbab 3.7 yang memakai pasangan ambang tetap, Tabel 7 melaporkan operating point hasil selection tanpa melihat outer-test. Seluruh baris memakai probabilitas akhir lima kelas dari eksperimen tiga partisi. Pada setiap fold, pasangan threshold dipilih hanya pada threshold-selection dengan objective coverage maksimum pada accepted accuracy minimal 95%, kemudian diterapkan satu kali pada outer-test. Confidence tetap 0,60; 0,70; 0,80; dan 0,90 dengan margin 0,20 merupakan grid demonstratif dalam kode, bukan kandidat yang dipilih dari skor outer-test untuk menentukan operating point final. Statusnya adalah pembandingan tetap, bukan bukti praregistrasi. Hanya baris "Dari threshold-selection" merupakan operating point hasil pemilihan pada partisi internal terpisah.

Tabel 7. Cakupan dan Reliabilitas Prediksi Selektif dengan Threshold-Selection Terpisah (Conf.=Confidence; Cov.=Coverage; Acc.=Accuracy)

Conf.	Margin	Cov.	Ditahan	Acc. diterima	Macro F1 diterima
Tanpa abstensi	0	100,00%	0	87,61%	88,17%
0,6	0,2	90,08%	745	91,53%	92,01%
0,7	0,2	83,23%	1.259	93,63%	93,96%
0,8	0,2	74,64%	1.904	95,77%	95,92%
0,9	0,2	62,20%	2.838	97,69%	97,69%
Dari threshold-selection	0	75,87%	1.812	95,40%	95,58%

Pada seluruh outer fold, ambang margin yang terpilih adalah 0,00, sedangkan ambang confidence yang terpilih memiliki nilai rata-rata 0,786 $\pm$ 0,036 (rentang 0,74–0,83). Karena margin probabilitas antara dua kelas dengan probabilitas tertinggi tidak bernilai negatif, kriteria margin $\geq$ 0,00 selalu terpenuhi. Dengan demikian, pada operating point yang terpilih, keputusan untuk menerima atau menolak prediksi secara efektif ditentukan oleh ambang confidence. Hasil ini menunjukkan bahwa penerapan ambang margin belum memberikan manfaat penyaringan tambahan pada konfigurasi terpilih. Oleh karena itu, hasil prediksi selektif diinterpretasikan berdasarkan operating point aktual tersebut, dengan coverage sebesar 75,87%, accepted accuracy sebesar 95,40%, dan accepted Macro F1 sebesar 95,58%. Sebanyak 5.696 keputusan diterima dan 1.812 ditahan. Nilai outer-test tersebut merupakan hasil pengamatan, bukan jaminan performa pada data baru. Pola reliabilitas divisualisasikan pada Gambar 3.



Gambar 3. Reliability Diagram Empat Metode Kalibrasi pada Protokol Tiga Partisi

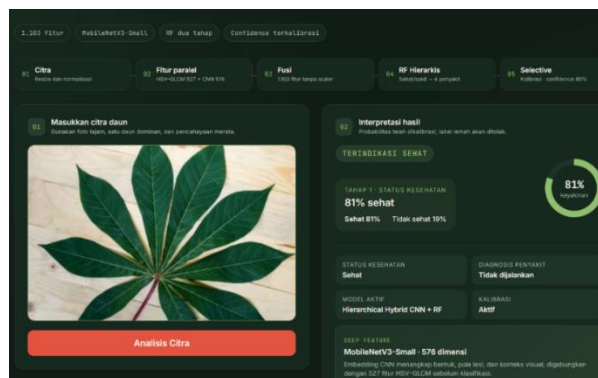
Pada panel tanpa kalibrasi di Gambar 3, sebagian besar bin confidence rendah hingga menengah berada di atas garis diagonal, yang menunjukkan kecenderungan under-confidence pada bin yang terisi. Interpretasi ini tidak digeneralisasi ke seluruh rentang confidence. Ketiga calibrator mendekatkan sebagian besar bin ke garis diagonal. Simplex-temperature menghasilkan ECE agregat terendah 0,54%, sedangkan sigmoid lebih baik pada log loss dan Brier; tidak ada satu calibrator yang unggul pada seluruh metrik.

3.10 Implementasi dan Pengujian Sistem

Implementasi web digunakan untuk memverifikasi konsistensi inferensi dengan pipeline eksperimen. Backend mengekstraksi 1.103 fitur tanpa scaler dan menjalankan dua Random Forest terkalibrasi. Antarmuka menerapkan gerbang Tahap 1 pada probabilitas Healthy-Diseased dan gerbang diagnosis pada probabilitas akhir lima kelas. Keluaran parsial 'tidak sehat' dapat dipertahankan ketika gerbang Tahap 1 lolos tetapi diagnosis final ditahan. Implementasi ini tidak diposisikan sebagai kontribusi algoritmik utama.

Gambar 4 menampilkan satu contoh citra sehat untuk mendemonstrasikan keputusan tahap kesehatan dan penghentian alur diagnosis penyakit. Screenshot memakai confidence 0,80 dan margin 0,20 sebagai konfigurasi demonstratif aplikasi, bukan operating point terpilih pada Tabel 7. Contoh ini tidak merupakan validasi numerik tambahan dan tidak menampilkan kasus citra tidak sehat.

Pada Gambar 4, classifier tahap pertama menghasilkan probabilitas sehat sebesar 81% dan tidak sehat sebesar 19%. Confidence 81% melampaui ambang 80% dan margin antarkelas sebesar 62 poin melampaui ambang 20 poin, sehingga keputusan sehat diterima. Karena citra diputuskan sehat, classifier tahap kedua untuk jenis penyakit tidak dijalankan.



Gambar 4. Demonstrasi Sistem pada Citra Daun Singkong Sehat dengan Ambang Tetap.

3.11 Pembahasan

Karakteristik CCMT berupa variasi latar belakang, pencahayaan, dan distribusi kelas [2] menunjukkan bahwa performa klasifikasi tidak cukup dinilai dari accuracy saja. Pada penelitian ini, Fusion RF Balanced meningkatkan balanced accuracy, tetapi menurunkan accuracy dan Macro F1 dibandingkan Fusion RF Standard. Oleh karena itu, evaluasi perlu mempertimbangkan keseimbangan antarkelas dan recall setiap kelas, bukan hanya satu metrik agregat.

Berbeda dari Zhong dkk. [9] yang menggunakan pendekatan deep learning berbasis Transformer-Embedded ResNet, penelitian ini tidak mengklaim model hierarkis Random Forest lebih unggul daripada deep learning end-to-end. MobileNetV3-Small memperoleh kinerja prediktif tertinggi, sedangkan Hierarchical Hybrid RF menawarkan pemisahan skrining sehat/sakit dan diagnosis penyakit. Kontribusi penelitian terletak pada integrasi fusi fitur, kalibrasi probabilitas, dan mekanisme menahan prediksi ketika keyakinan model belum memadai.

Hasil ablasi sejalan dengan penelitian fitur warna dan tekstur untuk klasifikasi penyakit tanaman [6] – [8]. HSV menjadi fitur handcrafted yang dominan, sementara enam fitur GLCM memberikan tambahan terbatas. Hasil ini menunjukkan bahwa kontribusi fitur bergantung pada karakteristik dataset dan kondisi citra. Selain itu, SHAP pada baseline handcrafted menunjukkan atribusi prediktif, bukan hubungan kausal dengan gejala biologis, karena fitur warna juga dapat dipengaruhi pencahayaan dan latar belakang.

Pada kalibrasi, simplex-temperature menghasilkan ECE terendah, sedangkan sigmoid lebih baik pada log loss dan Brier score. Temuan ini mendukung pentingnya mengevaluasi reliabilitas probabilitas menggunakan beberapa metrik [14]. Mekanisme prediksi selektif juga sejalan dengan gagasan hierarchical selective classification [17]. Sistem menerima 75,87% sampel dengan accepted accuracy 95,40%, sedangkan 24,13% prediksi ditahan. Dengan demikian, manfaat utama pendekatan ini adalah mengelola ketidakpastian dan meningkatkan ketepatan pada sampel yang diterima, bukan meningkatkan akurasi seluruh sampel.

4. KESIMPULAN

Penelitian ini mengevaluasi fusi fitur, klasifikasi hierarkis, kalibrasi probabilitas, dan prediksi selektif pada 7.508 citra CCMT. MobileNetV3-Small end-to-end memiliki rerata accuracy 93,90%, balanced accuracy 94,64%, dan Macro F1 94,23%, lebih tinggi daripada Hierarchical Hybrid RF pada perbandingan struktur. Hierarchical Hybrid menghasilkan rerata Macro F1 lebih tinggi secara deskriptif sebesar 1,03 poin persentase daripada Flat Hybrid, tetapi perbedaannya belum signifikan pada taraf 5% (Wilcoxon exact dua sisi, $W=1$; $p=0,125$), sehingga belum cukup bukti untuk menyimpulkan keunggulan model hierarkis. Pada protokol tiga partisi, simplex-temperature memiliki ECE terendah, sedangkan sigmoid lebih baik pada log loss dan Brier. Prediksi selektif menerima 75,87% sampel dengan accepted accuracy 95,40% dan accepted Macro F1 95,58%; margin terpilih 0,00 tidak menambah penyaringan. Manfaatnya adalah pengelolaan ketidakpastian dengan menahan 24,13% prediksi, bukan peningkatan akurasi intrinsik seluruh sampel atau jaminan performa baru. Tanpa metadata kelompok dan validasi eksternal, generalisasi ke tanaman atau lokasi baru belum dapat dipastikan.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: Regi Fernando Najoan, Enrico Samuel Djimesha, dan Argha Orion Silitonga;
Metodologi: Regi Fernando Najoan, Enrico Samuel Djimesha, dan Argha Orion Silitonga;
Perangkat Lunak: Regi Fernando Najoan dan Wilsen Grivin Mokodaser;
Validasi: Regi Fernando Najoan, Enrico Samuel Djimesha, dan Wilsen Grivin Mokodaser;
Analisis Formal: Regi Fernando Najoan dan Argha Orion Silitonga;
Investigasi: Regi Fernando Najoan, Enrico Samuel Djimesha, dan Argha Orion Silitonga;
Sumber Daya: George Morris William Tangka;
Kurasi Data: Regi Fernando Najoan dan Wilsen Grivin Mokodaser;
Penulisan Draf Awal: Regi Fernando Najoan;
Penulisan, Peninjauan, dan Penyuntingan: Regi Fernando Najoan, Enrico Samuel Djimesha, Argha Orion Silitonga, Wilsen Grivin Mokodaser, dan George Morris William Tangka;
Visualisasi: Regi Fernando Najoan dan Wilsen Grivin Mokodaser;
Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis korespondensi.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan khusus dari lembaga pemerintah, komersial, atau organisasi nirlaba.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan



Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.

REFERENCES

- [1] E. Mwebaze, T. Gebru, A. Frome, S. Nsumba, and J. Tusubira, “iCassava 2019 Fine-Grained Visual Categorization Challenge,” 2019. doi: 10.48550/arXiv.1908.02900.
- [2] P. K. Mensah et al., “CCMT: Dataset for Crop Pest and Disease Detection,” *Data Brief*, vol. 49, p. 109306, Aug. 2023, doi: 10.1016/j.dib.2023.109306.
- [3] V. K. Patel, K. Abhishek, and S. Selvarajan, “Optimized Recurrent Neural Network-Based Early Diagnosis of Crop Pest and Diseases in Agriculture,” *Discover Computing*, vol. 27, no. 1, p. 43, Nov. 2024, doi: 10.1007/s10791-024-09481-2.
- [4] B. Jia, Z. Zheng, L. Shi, H. Ge, and C. Song, “Interpretable Crop Pest and Disease Identification Based on Comparative Concept Tree,” *PLoS One*, vol. 21, no. 3, p. e0343715, Mar. 2026, doi: 10.1371/journal.pone.0343715.
- [5] A. M. Anggraeni, T. I. Hermanto, and I. M. Nugroho, “Klasifikasi Penyakit Daun Singkong Menggunakan Convolutional Neural Network (CNN) dengan Arsitektur VGG16 Berbasis Android,” *Jurnal Teknologi Terpadu*, vol. 11, no. 1, pp. 53–59, Jul. 2025, [Online]. Available: <https://journal.nurulfikri.ac.id/index.php/JTT/article/view/1862>
- [6] G. A. Defanka, A. Junaidi, and H. Maulana, “Penerapan LBP dan FCH untuk Klasifikasi Penyakit Daun Jagung Menggunakan Algoritma Random Forest,” *Jurnal Informatika Teknologi dan Sains (JINTEKS)*, vol. 8, no. 1, pp. 85–90, Feb. 2026, [Online]. Available: <https://jurnal.uts.ac.id/JINTEKS/article/view/6948>
- [7] A. A. Huda, B. Setiaji, and F. R. Hidayat, “Implementasi Gray Level Co-Occurrence Matrix (GLCM) untuk Klasifikasi Penyakit Daun Padi,” *Pseudocode*, vol. 9, no. 1, pp. 33–38, Feb. 2022, doi: 10.33369/pseudocode.9.1.33-38.
- [8] S. K. Wildah, A. Latif, A. Mustopa, Suharyanto, M. S. Maulana, and A. Sasongko, “Klasifikasi Penyakit Daun Kopi Menggunakan Kombinasi Haralick, Color Histogram dan Random Forest,” *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, vol. 11, no. 1, pp. 35–39, Jan. 2023, doi: 10.26418/justin.v11i1.60985.
- [9] Y. Zhong, B. Huang, and C. Tang, “Classification of Cassava Leaf Disease Based on a Non-Balanced Dataset Using Transformer-Embedded ResNet,” *Agriculture*, vol. 12, no. 9, p. 1360, Sep. 2022, doi: 10.3390/agriculture12091360.
- [10] D. Tran-Anh, Q. Nguyen Huu, T. Nguyen Thi Phuong, and Q. Dao Thi Thuy, “Leaf Disease Classification with Multiple-Model Deep Learning,” *Journal of Intelligent & Fuzzy Systems*, vol. 46, no. 1, pp. 2811–2823, Jan. 2024, doi: 10.3233/JIFS-235940.
- [11] A. Howard et al., “Searching for MobileNetV3,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.
- [12] A. Dosset et al., “Cassava Disease Detection Using a Lightweight Modified Soft Attention Network,” *Pest Manag. Sci.*, vol. 81, no. 2, pp. 607–617, Feb. 2025, doi: 10.1002/ps.8456.
- [13] H.-T. Thai, K.-H. Le, and N. L.-T. Nguyen, “FormerLeaf: An Efficient Vision Transformer for Cassava Leaf Disease Detection,” *Comput. Electron. Agric.*, vol. 204, p. 107518, Jan. 2023, doi: 10.1016/j.compag.2022.107518.
- [14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, PMLR, Sydney, Australia, 2017, pp. 1321–1330. [Online]. Available: <https://proceedings.mlr.press/v70/guo17a.html>
- [15] A. Niculescu-Mizil and R. Caruana, “Predicting Good Probabilities with Supervised Learning,” in *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, Bonn, Germany, 2005, pp. 625–632. doi: 10.1145/1102351.1102430.
- [16] Y. Geifman and R. El-Yaniv, “Selective Classification for Deep Neural Networks,” in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 4878–4887. [Online]. Available: <https://papers.nips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html>
- [17] S. Goren, I. Galil, and R. El-Yaniv, “Hierarchical Selective Classification,” in *Advances in Neural Information Processing Systems 37 (NeurIPS)*, Vancouver, Canada, 2024, pp. 111047–111073. doi: 10.52202/079017-3526.
- [18] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. New York, NY, USA: Pearson, 2018.
- [19] R. Dijaya, *Buku Ajar Pengolahan Citra Digital*. Sidoarjo: UMSIDA Press, 2023. doi: 10.21070/2023/978-623-464-075-5.
- [20] Scikit-learn Developers, “sklearn.ensemble.RandomForestClassifier,” 2025. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- [21] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 4765–4774. [Online]. Available: <https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>