Penerapan Ensemble CNN untuk Klasifikasi Biji Kopi: ResNet50, Inception V3, dan EfficientNet B7

Gabriel[1*], Wiiliam Parongko[2*], Aryo Michael[3*]

[1,2,3*]Program Studi Teknik Informatika, Universitas Kristen Indonesia Toraja, Tana Toraja, Sulawesi Selatan
Email: [1*]gkamma27@gmail.com

Abstrak

Indonesia, sebagai salah satu produsen dan eksportir kopi terbesar dunia, menghadapi tantangan dalam memenuhi standar kualitas global yang berpengaruh signifikan terhadap harga pasar. Proses penjaminan mutu dari budidaya hingga pascapanen memerlukan inovasi teknologi untuk meningkatkan efisiensi. Penelitian ini mengusulkan penerapan ensemble learning berbasis Convolutional Neural Network (CNN) dengan metode simple averaging yang mengkombinasikan tiga arsitektur pra-terlatih: ResNet-50, InceptionV3, dan EfficientNetB7. Dataset USK Coffee—terdiri dari 6.400 citra biji kopi berukuran 112×112 piksel—dibagi menjadi 4.800 sampel pelatihan dan 1.600 sampel pengujian untuk mengevaluasi kinerja model. Hasil eksperimen menunjukkan bahwa model ensemble mencapai akurasi 83%, mengungguli kinerja masing-masing model tunggal (ResNet-50: 77%, InceptionV3: 71%, EfficientNetB7: 82%). Analisis metrik precision, recall, dan F1-score mengonfirmasi peningkatan signifikan dalam konsistensi klasifikasi, khususnya pada kelas defect (precision 0.90) dan longberry (F1-score 0.91). Namun, disparitas kinerja pada kelas peaberry (precision 0.74 vs. recall 0.93) mengindikasikan perlunya optimasi tambahan melalui augmentasi data atau penyesuaian threshold. Temuan ini menegaskan potensi teknik ensemble dalam sistem klasifikasi biji kopi berbasis deep learning, sekaligus menyoroti kebutuhan penanganan kompleksitas komputasi dan ambiguitas fitur visual untuk aplikasi industri skala besar.

Kata Kunci: Ensemble Learning, Convolutional Neural Network, Klasifikasi Biji Kopi, Deep Learning, Transfer Learning.

Implementation of Ensemble CNN for Coffee Bean Classification: ResNet50, Inception V3, and EfficientNet B7

Abstract

As one of the world's leading coffee producers and exporters, Indonesia faces challenges in meeting global quality standards, which significantly influence market pricing. Ensuring quality from cultivation to post-harvest processing requires technological innovation to enhance efficiency. This study proposes a Convolutional Neural Network (CNN)-based ensemble learning approach using a simple averaging technique that combines three pre-trained architectures: ResNet-50, InceptionV3, and EfficientNetB7. The USK Coffee dataset—comprising 6,400 coffee bean images (112×112 pixels)—was divided into 4,800 training and 1,600 testing samples to evaluate model performance. Experimental results demonstrate that the ensemble model achieved 83% accuracy, outperforming individual models (ResNet-50: 77%, InceptionV3: 71%, EfficientNetB7: 82%). Analysis of precision, recall, and F1-score metrics confirmed

significant improvements in classification consistency, particularly for the defect class (precision: 0.90) and longberry class (F1-score: 0.91). However, performance disparities in the peaberry class (precision: 0.74 vs. recall: 0.93) indicate the need for further optimization through data augmentation or threshold adjustment. These findings validate the potential of ensemble techniques in coffee bean classification systems while highlighting the necessity to address computational complexity and visual feature ambiguity for large-scale industrial applications.

Keywords: Ensemble Learning, Convolutional Neural Network, Coffee Bean Classification, Deep Learning, Transfer Learning.

I. PENDAHULUAN

Indonesia merupakan salah satu negara produsen dan eksportir kopi terbesar di dunia, dengan kontribusi signifikan terhadap pasar global [1]. Perkembangan industri kopi mengalami pertumbuhan dinamis di seluruh rantai pasok, mulai dari tingkat budidaya oleh petani, distribusi pemasok, bisnis ritel *café*, hingga konsumsi akhir [2]. Harga komoditas kopi internasional sangat bergantung pada parameter kualitas, yang pencapaiannya memerlukan proses terintegrasi mulai dari fase penanaman hingga pengolahan pascapanen [3].

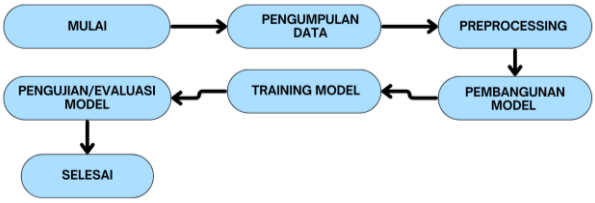
Salah satu tahapan kritis dalam pascapanen adalah grading (klasifikasi biji kopi) berdasarkan standar mutu yang ditetapkan [4]. Di pasar global, kopi umumnya diperdagangkan dalam bentuk *green bean* (biji hijau), di mana kualitasnya ditentukan melalui seleksi berdasarkan atribut warna, cacat fisik, dan varietas biji [5]. Proses klasifikasi yang masih mengandalkan metode manual menghadapi tantangan efisiensi, seperti durasi proses yang panjang dan inkonsistensi hasil sortasi, sehingga berpotensi menurunkan nilai jual produk [6].

Dalam konteks revolusi industri 4.0, teknologi *computer vision* yang diintegrasikan dengan algoritma kecerdasan buatan (AI) menjadi solusi inovatif untuk otomasi sortasi. *Convolutional Neural Network* (CNN) merupakan algoritma *deep learning* yang banyak diaplikasikan dalam klasifikasi citra karena kemampuannya mengekstraksi fitur hierarkis [7]. Namun, implementasi CNN memiliki beberapa keterbatasan: Kebutuhan Dataset Besar: Pelatihan model CNN rentan terhadap *vanishing gradient*, yang dapat diatasi dengan arsitektur residual seperti ResNet [8]. Kompleksitas Komputasi: Waktu pelatihan yang lama memerlukan optimasi arsitektur, misalnya melalui desain *Inception* atau *EfficientNet*. Stokastisitas Prediksi: Variabilitas hasil klasifikasi akibat sifat probabilistik jaringan saraf memerlukan pendekatan *ensemble learning* untuk meningkatkan stabilitas.

Berdasarkan latar belakang tersebut, penelitian ini difokuskan pada eksplorasi kinerja model *ensemble* yang mengkombinasikan arsitektur ResNet-50, Inception V3, dan EfficientNet B7 untuk klasifikasi biji kopi.

II. METODE PENELITIAN

A. Tahapan Penelitian



Gambar 1. Tahapan Penelitian

1) Pengumpulan Data (*Data Collection*)

Tahap awal penelitian melibatkan akuisisi dataset citra biji kopi hijau (*green bean*) dari sumber terpercaya, yaitu USK-Coffee Dataset. Dataset ini terdiri atas empat kelas morfologis biji kopi arabika, yaitu *Longberry*, *Peaberry*, *Premium*, dan *Defect*, dengan total 6.400 citra beresolusi seragam. Contoh gambar setiap kelas disajikan pada Gambar 2.

Gambar 2. (a) *longberry*, (b) *peaberry*,(c) *premium*, (d) defect.

Distribusi data dibagi menjadi dua subset: Subset Pelatihan: 4.800 citra (1.200 citra per kelas). Subset Pengujian: 1.600 citra (400 citra per kelas). Seluruh data kemudian diunggah ke platform penyimpanan cloud (*Google Drive*) dan diproses menggunakan *Google Colaboratory* untuk memanfaatkan sumber daya komputasi berbasis GPU secara efisien.

2) Preprocessing

Preprocessing merupakan tahap kritis untuk mempersiapkan data agar kompatibel dengan arsitektur model. Proses ini meliputi tiga langkah utama:

1. Resizing Citra

Proses ini bertujuan untuk menyeragamkan dimensi spasial citra dengan mengubah ukuran seluruh citra menjadi 112×112 piksel. Penyesuaian resolusi dilakukan untuk mengoptimasi efisiensi komputasi, mengingat ukuran citra yang besar memerlukan kapasitas memori tinggi dan memperpanjang durasi pelatihan. Reduksi dimensi citra secara signifikan mengurangi kompleksitas komputasi tanpa mengorbankan fitur esensial biji kopi.

2. Labeling

Labeling bertransformasi dari format kategorikal ke representasi numerik melalui teknik integer encoding. Konversi ini memfasilitasi pemrosesan komputasi dan memastikan kompatibilitas dengan lapisan *output* model yang menggunakan aktivasi *softmax*.

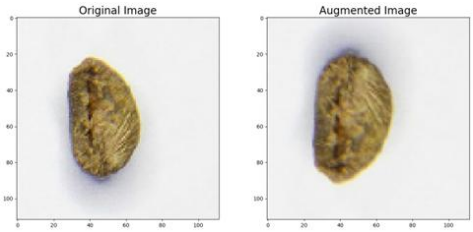
3. Augmentasi Data

Augmentasi data diterapkan untuk meningkatkan variasi dataset tanpa penambahan sampel baru, sehingga mencegah *overfitting*. Dua teknik utama yang digunakan:

- Pembalikan Acak (Random Flip): Transformasi acak pada sumbu horizontal (X) dan vertikal (Y) untuk meningkatkan ketahanan model terhadap variasi orientasi citra.
- Zoom Acak (Random Zoom): Penskalaan citra dengan faktor ±10% (rentang: [-0.1, 0.1]) untuk mengekspos model pada variasi skala objek. Teknik ini mendorong model belajar fitur invariant-skala, yang kritis dalam generalisasi ke data baru.

Proses augmentasi diimplementasikan secara dinamis selama pelatihan menggunakan lapisan *ImageDataGenerator* pada *TensorFlow/Keras*.

```
# Build augmentation layer
augmentation_layer = Sequential([
    layers.RandomFlip(mode='horizontal_and_vertical', seed=CFG.TF_SEED),
    layers.RandomZoom(height_factor=(-0.1, 0.1), width_factor=(-0.1, 0.1), seed=CFG.TF_SEED),
], name='augmentation_layer')
```



Gambar 3. Potongan kode dan Hasil Preprocessing

3). Pembangunan Model

Pada tahapan ini dibangun model CNN, Pembangunan model CNN menggunakan pendekatan transfer learning. Arsitektur ResNet50, InceptionV3 dan EfficienNetB7 dibangun secara terpisah, selanjutnya model ensemble learning digunakan untuk menggabungkan hasil prediksi dari setiap arsitektur CNN yang telah dibuat sebelumnya. Dalam Pembangunan model dengan transfer learning hal pertama yang dilakukan adalah mengunduh arsitektur ResNet50, InceptionV3 dan EfficienNetB7 yang telah dilatih menggunakan dataset imageNet selanjutnya dilakukan penyesuaian sesuai dengan jumlah kelas biji kopi yang diklasifikasikan.

4). Training Model

Setelah tahapan pembangunan model *Convolutional Neural Network* (CNN), proses pelatihan model tunggal (ResNet-50, InceptionV3, dan EfficientNetB7) dilakukan untuk melatih kemampuan masing-masing arsitektur dalam mengidentifikasi kelas biji kopi. Pelatihan model dilaksanakan menggunakan *dataset* pelatihan yang dibagi menjadi dua subset dengan rasio 80:20, di mana 80% (3.840 data) dialokasikan sebagai *training set* dan 20% (960 data) berfungsi sebagai *validation set*. Parameter pelatihan ditetapkan dengan kombinasi 50 *epoch* dan *batch size* 32 untuk mengoptimalkan proses pembelajaran. Dalam implementasinya, dua *callback function* diterapkan untuk meningkatkan efisiensi pelatihan, yakni *EarlyStopping* (untuk menghentikan pelatihan secara otomatis ketika metrik validasi tidak menunjukkan peningkatan) dan *ReduceLROnPlateau* (untuk menurunkan *learning rate* secara dinamis ketika akurasi validasi stagnan). Kedua mekanisme ini berperan dalam mengontrol konsistensi pelatihan guna mencegah *overfitting* dan memaksimalkan generalisasi model.

4). Pengujian dan Evaluasi Model

Tahap akhir dalam siklus pengembangan model adalah **evaluasi performa**, yang bertujuan mengukur kemampuan model dalam melakukan klasifikasi berdasarkan data uji. Salah satu metode evaluasi utama adalah *confusion matrix* (matriks konfusi), yang memvisualisasikan distribusi hasil prediksi model terhadap label aktual melalui empat parameter kuantitatif: *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN) [9]. Berdasarkan parameter tersebut, metrik evaluasi seperti akurasi (*accuracy*), presisi (*precision*), *recall*, dan *F1-score* dihitung untuk mengkuantifikasi performa model secara komprehensif (Persamaan 1-4). Hasil perhitungan *confusion matrix* pada penelitian ini disajikan dalam Tabel 2, sementara formula matematis setiap metrik mengacu pada persamaan yang telah ditetapkan sebelumnya. Analisis terhadap metrik ini tidak hanya mengidentifikasi keunggulan model dalam mengklasifikasikan kelas tertentu, tetapi juga mengungkap kelemahan seperti bias prediksi atau ketidakseimbangan sensitivitas antar-kelas, sehingga menjadi dasar rekomendasi perbaikan model.

$$Accurasy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

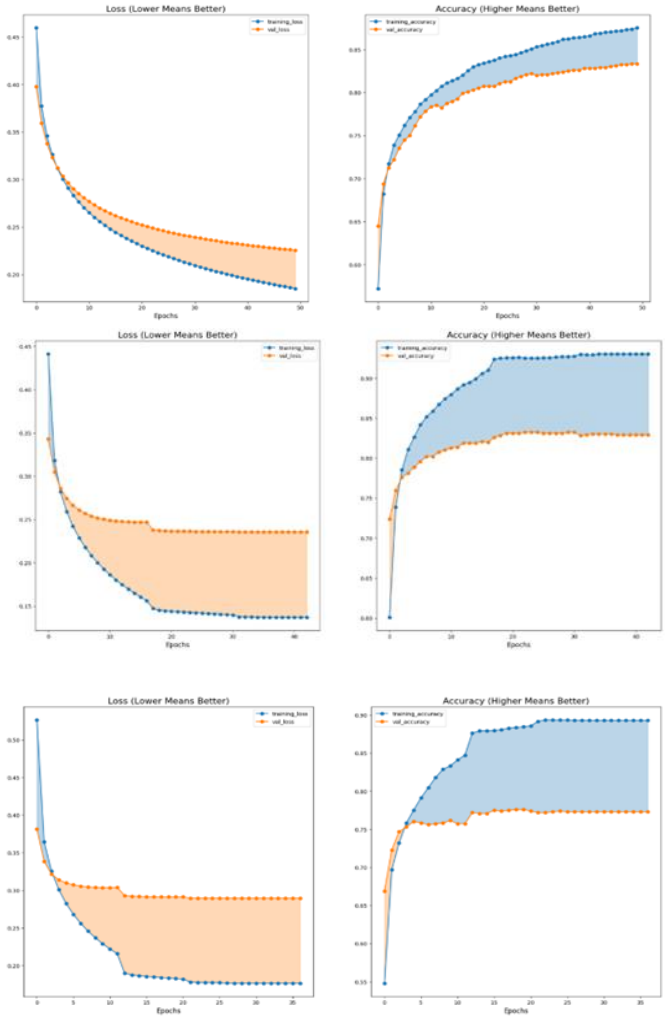
$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$f1 - score = 2\left(\frac{Precision\ x\ Recall}{Precision + Recall}\right) \quad (4)$$

III. HASIL DAN PEMBAHASAN

. Setelah menyelesaikan fase pelatihan, model yang telah terlatih (*trained model*) siap diuji menggunakan data independen untuk menilai kinerja generalisasinya. Pada fase pengujian, model menerima input data baru dan menghasilkan prediksi berdasarkan pola yang dipelajari selama pelatihan. Proses ini memungkinkan evaluasi terhadap kemampuan generalisasi model dalam memproses sampel yang tidak tercakup dalam *training set*. Selama pelatihan, dua metrik utama termonitor secara iteratif: nilai akurasi yang merepresentasikan persentase prediksi benar terhadap total sampel, dan nilai loss yang mengkuantifikasi besarnya deviasi antara prediksi model dengan nilai sebenarnya (*ground truth*). Optimasi model bertujuan meminimalkan fungsi *loss* sembari memaksimalkan akurasi, di mana nilai *loss* yang rendah menunjukkan konvergensi model mendekati solusi optimal, sedangkan akurasi tinggi mencerminkan konsistensi prediksi sesuai label aktual. Tujuan akhir pembangunan model adalah memperoleh konfigurasi bobot (*weights*) yang mencapai *minimum global* pada fungsi *loss* melalui penyesuaian parameter selama pelatihan, sehingga memastikan stabilitas dan keandalan prediksi pada data uji.

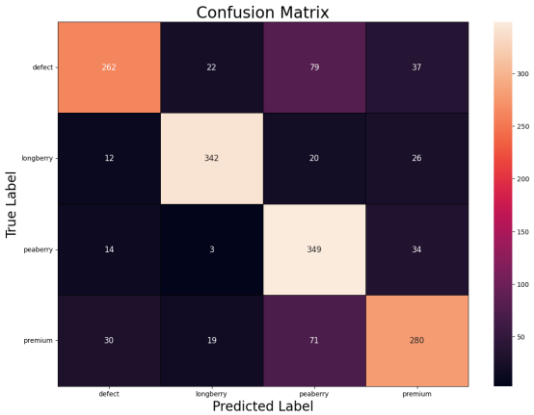


Gambar 13 Grafik Training Model EfficienNetB7, ResNet-50 dan Inception V3

Berdasarkan hasil yang ditampilkan pada Gambar 13, proses pelatihan (*training*) pada ketiga arsitektur model menunjukkan pola terminasi yang berbeda. Model berbasis arsitektur EfficientNetB7 menyelesaikan seluruh iterasi pelatihan hingga *epoch* ke-50, sedangkan arsitektur ResNet-50 mengalami terminasi dini pada *epoch* ke-43, dan InceptionV3 berhenti pada *epoch* ke-37. Variasi waktu terminasi ini disebabkan oleh implementasi mekanisme *callback* EarlyStopping, yang menghentikan pelatihan secara otomatis ketika metrik validasi (misalnya, akurasi atau *loss*) tidak menunjukkan peningkatan signifikan dalam rentang *epoch* tertentu. Hasil ini mengindikasikan bahwa InceptionV3 mencapai konvergensi lebih cepat dibandingkan dua arsitektur lain, sementara EfficientNetB7 memerlukan iterasi lebih panjang untuk mengekstraksi fitur kompleks akibat kedalaman arsitekturnya. Perbedaan ini merefleksikan karakteristik unik setiap arsitektur dalam adaptasi terhadap dataset pelatihan serta efisiensi optimasi bobot model.

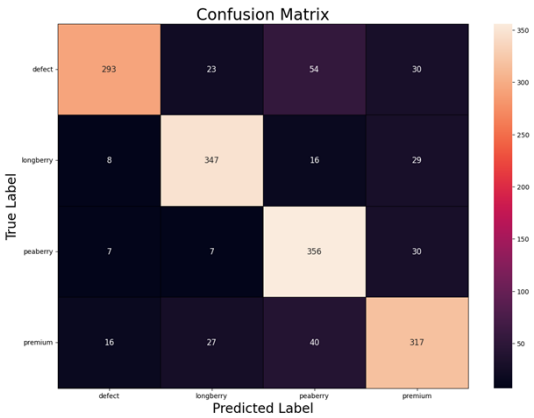
5). Pengujian Dan Evaluasi Model

Setelah menyelesaikan fase pelatihan model, tahap selanjutnya dilakukan evaluasi performa terhadap model tunggal (*single model*) dan model *ensemble*. Hasil evaluasi tersebut divisualisasikan melalui *confusion matrix* (matriks konfusi) untuk mengukur akurasi klasifikasi, presisi, *recall*, dan *F1-score* pada data uji. Distribusi hasil prediksi dari masing-masing model ditampilkan secara komparatif pada Gambar 14 hingga Gambar 17, yang mencakup analisis kinerja arsitektur individu (ResNet-50, InceptionV3, EfficientNetB7) serta kombinasi *ensemble*. Visualisasi ini memungkinkan identifikasi pola kesalahan klasifikasi, seperti ketidakseimbangan prediksi antar-kelas atau bias model tertentu, sekaligus membuktikan efektivitas pendekatan *ensemble* dalam meningkatkan stabilitas prediksi. Hasil tersebut menjadi dasar kuantitatif untuk menilai generalisasi model dan validasi hipotesis penelitian terkait optimasi arsitektur CNN pada klasifikasi biji kopi.



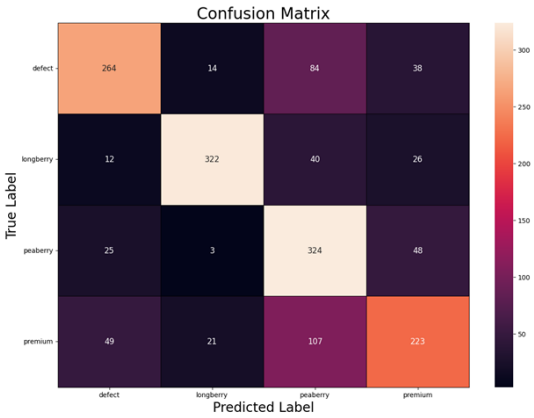
Gambar 14 Confusion Matrix ResNet-50

Berdasarkan Gambar 14, model tunggal ResNet-50 menunjukkan adanya kesalahan klasifikasi pada dua kelas spesifik: (1) sebanyak 37 sampel dari kelas *defect* teridentifikasi secara keliru sebagai kelas *premium*, dan (2) 26 sampel dari kelas *longberry* salah diprediksi sebagai kelas *premium*.



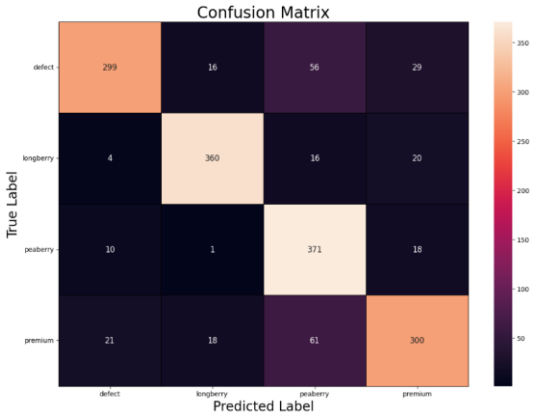
Gambar 15 Confusion Matrix EfficientNetB7

Berdasarkan Gambar 15, model tunggal EfficientNetB7 menunjukkan adanya kesalahan klasifikasi pada dua kasus spesifik: (1) sebanyak 16 sampel dari kelas *longberry* terdeteksi secara keliru sebagai kelas *peaberry*, dan (2) 27 sampel dari kelas *premium* salah diklasifikasikan ke dalam kelas *longberry*.



Gambar 16 Confusion Matrix Inception V3

Berdasarkan Gambar 16, model unggal InceptionV3 menunjukkan adanya kesalahan klasifikasi pada dua kategori: (1) 14 sampel dari kelas *defect* (cacat) terdeteksi secara keliru sebagai kelas *longberry*, dan (2) 25 sampel dari kelas *peaberry* salah diklasifikasikan ke dalam kelas *defect*.



Gambar 17 Confusion Matrix Simple Average Ensemble

Berdasarkan Gambar 17, model *ensemble* dengan metode *simple averaging* masih menunjukkan kesalahan klasifikasi pada dua kasus: (1) sebanyak 20 sampel dari kelas *longberry* terdeteksi secara keliru sebagai kelas *premium*, dan (2) 16 sampel dari kelas *defect* (cacat) salah diklasifikasikan ke dalam kelas *longberry*.

Setelah dilakukan pengujian, selanjutnya dilakukan evaluasi menggunakan melihat nilai akurasi, presisi, recall dan f1-score masing-masing model.

1). Model ResNet-50

Berdasarkan hasil evaluasi kuantitatif pada Gambar 18, model ResNet-50 mencapai akurasi sebesar 0,77 (77%) pada

dataset uji yang terdiri dari 1.600 sampel. Analisis metrik klasifikasi per kelas menunjukkan variasi performa: kelas *longberry* mencatat nilai *precision* tertinggi (0,89) dan *F1-score* optimal (0,87), mengindikasikan kemampuan model dalam mengidentifikasi kelas ini secara konsisten. Namun, kelas *peaberry* memiliki *recall* tertinggi (0,87) dengan *precision* relatif rendah (0,67), yang merefleksikan kecenderungan model menghasilkan *false positive* pada kategori ini. Sementara itu, kelas *defect* dan *premium* menunjukkan *F1-score* moderat (masing-masing 0,73 dan 0,72), menandakan adanya ambiguitas fitur atau tumpang tindih karakteristik visual antar-kelas tersebut. *Macro average* dan *weighted average* untuk *F1-score* (0,77) mengonfirmasi konsistensi performa model secara keseluruhan, meskipun disparitas antar-kelas masih perlu diatasi. Hasil ini menggarisbawahi perlunya optimasi tambahan, seperti augmentasi data selektif atau penyesuaian *decision threshold*, untuk meningkatkan diskriminasi fitur pada kelas dengan *precision* dan *recall* tidak seimbang.

	precision	recall	f1-score	support
defect	0.82	0.66	0.73	400
longberry	0.89	0.85	0.87	400
peaberry	0.67	0.87	0.76	400
premium	0.74	0.70	0.72	400
accuracy			0.77	1600
macro avg	0.78	0.77	0.77	1600
weighted avg	0.78	0.77	0.77	1600

Gambar 18 Report Kinerja Model ResNet50

2). Model EfficientNetB7

Berdasarkan hasil evaluasi kuantitatif pada Gambar 19, model EfficientNetB7 mencapai akurasi sebesar 0,82 (82%) pada dataset uji yang terdiri dari 1.600 sampel. Analisis metrik klasifikasi per kelas menunjukkan performa yang heterogen: kelas *defect* mencatat *precision* tertinggi (0,90), tetapi *recall* relatif rendah (0,73), mengindikasikan kecenderungan model menghindari *false positive* namun kurang sensitif dalam mendeteksi sampel cacat aktual. Sebaliknya, kelas *peaberry* mencapai *recall* tertinggi (0,89) dengan *precision* moderat (0,76), yang merefleksikan kemampuan model mendeteksi sebagian besar sampel *peaberry* meskipun menghasilkan beberapa *false positive*. Kelas *longberry* menunjukkan keseimbangan optimal antara *precision* (0,86) dan *recall* (0,87), menghasilkan *F1-score* tertinggi (0,86), sementara kelas *premium* memiliki kinerja stabil dengan *F1-score* 0,79. *Macro average* dan *weighted average* untuk *F1-score* (0,82) mengonfirmasi konsistensi model secara keseluruhan, meskipun disparitas kinerja antar-kelas, khususnya pada *defect* dan *peaberry*

	precision	recall	f1-score	support
defect	0.90	0.73	0.81	400
longberry	0.86	0.87	0.86	400
peaberry	0.76	0.89	0.82	400
premium	0.78	0.79	0.79	400
accuracy			0.82	1600
macro avg	0.83	0.82	0.82	1600
weighted avg	0.83	0.82	0.82	1600

Gambar 19 Report Kinerja Model EfficienNetB7

3). Model Inception V3

Berdasarkan hasil evaluasi kuantitatif pada Gambar 20, model InceptionV3 mencapai akurasi sebesar 0,71 (71%) pada dataset uji yang terdiri dari 1.600 sampel. Analisis metrik klasifikasi per kelas mengungkap disparitas kinerja yang signifikan: kelas longberry mencatat performa terbaik dengan precision 0,89, recall 0,81, dan F1-score 0,85, mengindikasikan kemampuan model dalam mengidentifikasi kelas ini secara stabil dengan minim kesalahan (false positive/negative). Sebaliknya, kelas defect menunjukkan ketidakseimbangan antara precision (0,75) dan recall (0,66), yang merefleksikan kecenderungan model menghasilkan false negative pada kategori cacat. Kelas peaberry memiliki recall tertinggi (0,81) namun precision terendah (0,58), menandakan dominasi false positive akibat ambiguitas fitur atau ketidakterwakilan sampel dalam pelatihan. Sementara itu, kelas premium mencatat F1-score terendah (0,61) dengan precision 0,67 dan recall 0,56, mengisyaratkan kesulitan model dalam membedakan karakteristik visual kelas ini dari kategori lain.

	precision	recall	f1-score	support
defect	0.75	0.66	0.70	400
longberry	0.89	0.81	0.85	400
peaberry	0.58	0.81	0.68	400
premium	0.67	0.56	0.61	400
accuracy			0.71	1600
macro avg	0.72	0.71	0.71	1600
weighted avg	0.72	0.71	0.71	1600

Gambar 20 Report Kinerja Model Inception V3

4). Model Simple Average Ensemble

Berdasarkan hasil evaluasi kuantitatif pada Gambar 21, model Simple Average Ensemble mencapai akurasi sebesar 0,83 (83%) pada dataset uji yang terdiri dari 1.600 sampel. Analisis metrik klasifikasi per kelas menunjukkan bahwa kelas longberry mencatat kinerja optimal dengan precision 0,91, recall 0,90, dan F1-score 0,91, mengindikasikan konsistensi model dalam mengidentifikasi kelas ini dengan minim kesalahan (false positive/negative). Kelas defect juga menunjukkan precision tinggi (0,90), namun recall relatif rendah (0,75), yang merefleksikan kecenderungan model menghasilkan false negative pada kategori cacat. Di sisi lain, kelas peaberry memiliki recall tertinggi (0,93) tetapi precision

terendah (0,74), menandakan dominasi false positive akibat ambiguitas fitur atau ketidakterwakilan sampel dalam pelatihan. Sementara itu, kelas premium mencatat F1-score 0,78 dengan precision 0,82 dan recall 0,75, mengisyaratkan kesulitan model dalam membedakan karakteristik visual kelas ini dari kategori lain.

	precision	recall	f1-score	support
defect	0.90	0.75	0.81	400
longberry	0.91	0.90	0.91	400
peaberry	0.74	0.93	0.82	400
premium	0.82	0.75	0.78	400
accuracy			0.83	1600
macro avg	0.84	0.83	0.83	1600
weighted avg	0.84	0.83	0.83	1600

Gambar 21 Report Kinerja Model Simple Average Ensemble

Gambar 22 merupakan grafik perbandingan kinerja setiap model, dimana terlihat bahwa kinerja model average ensemble mengungguli kinerja model CNN Tunggal. Ini berarti average ensemble akan menggeneralisasi ke sampel yang tidak terlihat karena MCC yang tinggi menyiratkan bahwa prediksi model secara statistik berkualitas tinggi.

Dari hasil pengujian yang dilakukan, selanjutnya dilakukan evaluasi kinerja setiap model menggunakan matriks Kinerja model didasarkan pada nilai accuracy score, precision, recall, f1-score dan matthew coefficient corelation. Adapun kinerja setiap model disajikan pada Tabel 1

Tabel 1 Pengukuran Kinerja Model

Model	*Accuracy*	*Precision*	*Recall*	*F1_Score*	*MCC*
EfficientNet B7	0.820625	0.826992	0.820625	0.820337	0.762859
Inception V3	0.708125	0.724546	0.708125	0.709176	0.616045
ResNet50	0.770625	0.781266	0.770625	0.770069	0.698106
Simple Average Ensemble	0.831250	0.840038	0.831250	0.830860	0.778308

Berdasarkan evaluasi kinerja empat model—ResNet-50, InceptionV3, EfficientNetB7, dan Simple Average Ensemble—model ensemble berbasis simple averaging mencatat akurasi tertinggi sebesar 83%, mengungguli model tunggal seperti ResNet-50 (77%), InceptionV3 (71%), dan EfficientNetB7 (82%). Keunggulan ensemble ini juga tercermin dari nilai macro dan weighted average F1-score (0,83), yang menunjukkan konsistensi kinerja agregat dibandingkan model tunggal. Secara spesifik, ensemble berhasil meningkatkan presisi kelas defect hingga 0,90—lebih tinggi daripada ResNet-50 (0,75) dan EfficientNetB7 (0,82)—meskipun recall-nya masih terbatas (0,75), mengindikasikan kecenderungan false negative. Pada kelas longberry, ensemble mempertahankan kinerja optimal dengan F1-score 0,91, sejalan dengan performa model tunggal terbaik (EfficientNetB7: 0,87). Namun, pada kelas peaberry, meskipun recall mencapai 0,93, precision yang rendah (0,74) menandakan dominasi false positive akibat ambiguitas fitur atau ketidakseimbangan data.

Kinerja ensemble dalam mengurangi varians prediksi dan menggabungkan kekuatan model tunggal terbukti efektif, terutama dalam meningkatkan akurasi holistik dan stabilitas klasifikasi. Namun, disparitas kinerja antar-kelas—seperti trade-off presisi-recall pada peaberry dan keterbatasan recall pada defect—menyoroti perlunya optimasi tambahan. Kompleksitas komputasional dan ketergantungan pada kualitas prediksi model dasar menjadi tantangan dalam pendekatan ensemble. Secara keseluruhan, metode ini valid sebagai strategi peningkatan generalisasi model, tetapi memerlukan augmentasi data selektif, penyesuaian threshold, atau integrasi meta-learner untuk mengatasi kelemahan pada kelas dengan variasi visual tinggi.

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Model ensemble yang menggabungkan ResNet-50, InceptionV3, dan EfficientNetB7 mencapai akurasi tertinggi (83%), mengungguli model tunggal (ResNet-50: 77%, InceptionV3: 71%, EfficientNetB7: 82%). Teknik ensemble meningkatkan generalisasi dengan mengurangi varians prediksi, terlihat dari peningkatan precision kelas defect (0,90) dan stabilitas recall kelas longberry (0,90). Namun, tantangan seperti trade-off precision-recall pada kelas peaberry (precision 0,74 vs. recall 0,93) serta kompleksitas komputasi perlu diatasi melalui optimasi threshold, augmentasi data, atau metode weighted averaging. Hasil ini membuktikan efektivitas ensemble learning dalam klasifikasi biji kopi berbasis CNN.

DAFTAR PUSTAKA

[1] T. Theodoridis and J. Kraemer, “indonesia,” pp. 1–9.
[2] A. Rivalto, “Klasifikasi Jenis Kopi Indonesia Menggunakan Deep Learning,” Undergrad. Thesis, vol. BAB I, 2019.
[3] Alnopri, “Upaya pengembangan inovasi teknologi untuk meningkatkan produksi dan mutu serta varian minuman kopi,” J. Agric. Nat. Resour., vol. 04, pp. 0–5, 2023, doi: 10.32734/anr.v4i1.1736.
[4] A. Michael and J. Rusman, “Klasifikasi Cacat Biji Kopi Menggunakan Metode Transfer Learning dengan Hyperparameter Tuning Gridsearch,” J. Teknol. dan Manaj. Inform., vol. 9, no. 1, pp. 37–45, 2023, doi: 10.26905/jtmi.v9i1.10035.
[5] Y. A. Musika, “Apa Itu Green Bean Kopi, Seperti Apa Manfaatnya?,” Otten, pp. 1–9, 2022, [Online]. Available: https://ottencoffee.co.id/majalah/green-bean-kopi
[6] C. Pramono, K. Suharno, and R. A. Putranto, “Pengaruh Waktu Grading Terhadap Kualitas Biji Kopi Arabika,” Semin. Nas. Edusaintek FMIPA UNIMUS, pp. 101–107, 2018.
[7] rahayu deny danar dan alvi furwanti Alwie, A. B. Prasetio, R. Andespa, P. N. Lhokseumawe, and K. Pengantar, “Tugas Akhir Tugas Akhir,” J. Ekon. Vol. 18, Nomor 1 Maret201, vol. 2, no. 1, pp. 41–49, 2020.

[8] E. A. Oktaviari, “Bab II Landasan Teori,” J. Chem. Inf. Model., vol. 53, no. 9, p. 1689, 2019, [Online]. Available: https://repository.bsi.ac.id/index.php/unduh/item/257726/File-10-BAB-II.pdf

[9] A. Kurniadi, “Implementasi Convolutional Neural Network Untuk Klasifikasi Varietas Pada Citra Daun Sawi Menggunakan Keras,” Doubleclick, vol. 4, no. 1, hlm. 25, Ag

.