Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.20241117389
Vol. 11, No. 1, Februari 2024, hlm. 75-84 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

VISION TRANSFORMER UNTUK KLASIFIKASI KEMATANGAN PISANG

Arya Pangestu[1], Bedy Purnama*[2], Risnandar[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
Email: [1]aryapangestu@student.telkomuniversity.ac.id, [2]bedypurnama@telkomuniversity.ac.id,
[3]risnandartelyu@telkomuniversity.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 21 Juni 2023, diterima untuk diterbitkan: 20 November 2023)

Abstrak

Produksi pisang di Indonesia pada tahun 2022 mencapai 9,6 juta ton buah. Metode konvensional yang digunakan untuk menentukan tingkat kematangan pisang masih mengandalkan indera penglihatan manusia dengan memperhatikan perubahan warna kulit pisang. Namun, penentuan tingkat kematangan pisang dengan metode ini memiliki beberapa kekurangan, seperti waktu yang lama, penilaian yang bersifat subjektif dan dapat menghasilkan hasil yang berbeda-beda bagi setiap individu. Oleh karena itu, teknologi *computer vision* dapat menjadi solusi yang efektif dalam mengklasifikasikan kematangan buah pisang secara otomatis. Penelitian ini menggunakan metodologi Vision Transformer (ViT) untuk mengklasifikasikan tingkat kematangan pada buah pisang, dengan tingkatan yang dibagi menjadi empat kategori, yaitu mentah, setengah matang, matang, dan terlalu matang. Penelitian dilakukan dengan menggunakan lima model ViT yang sudah dilatih sebelumnya atau *pre-trained*, yaitu ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32, and ViT-H/14 pada ImageNet-21k dan ImageNet-1k. Kemudian, model ViT tersebut dievaluasi dan dibandingkan dengan model CNN. Evaluasi dilakukan menggunakan metode *cross-dataset* dengan 5.068 citra pisang yang berbeda dari dataset latih. Hasil evaluasi menunjukkan model ViT-L/16-in21k memiliki akurasi tertinggi sebesar 91,61%. Model ViT menunjukkan kemampuan generalisasi yang lebih baik, sementara CNN memiliki ukuran model dan waktu pelatihan yang lebih efisien.

Kata kunci: *klasifikasi, kematangan pisang, computer vision, vision transformer, pre-trained model, cross-dataset evaluation*

VISION TRANSFORMER FOR BANANA RIPENESS CLASSIFICATION

Abstract

Banana production in Indonesia in 2022 reached 9.6 million tons of fruit. The conventional method used to determine the ripeness level of bananas still relies on human sight by observing changes in the color of the banana skin. However, determining the ripeness level of bananas using this method has disadvantages, such as taking a long time, subjective assessment, and producing different results for each individual. Therefore, computer vision technology can be an effective solution in automatically classifying the ripeness level of bananas. This study uses Vision Transformer (ViT) methodology to classify the ripeness level of bananas into four categories: unripe, half-ripe, ripe, and overripe. The study used five pre-trained ViT models, namely ViT-B/16, ViT- B/32, ViT-L/16, ViT-L/32, and ViT-H/14 on ImageNet-21k and ImageNet-1k. Then, the ViT models were evaluated and compared with CNN models. The evaluation is conducted using the cross-dataset method with 5,068 different banana images from the training dataset. The results show that the ViT-L/16-in21k model achieves the highest accuracy at 91.61%. ViT models demonstrate better generalization abilities, while CNN models exhibit more efficient model size and training time.

Keywords: *classification, banana ripeness, computer vision, vision transformer, pre-trained model, cross-dataset evaluation*

1. PENDAHULUAN

Produksi pisang di Indonesia meningkat dari 8,7 juta ton pada tahun 2021 menjadi 9,6 juta ton pada tahun 2022 (Indonesia, 2023). Pisang mengandung berbagai vitamin dan mineral yang memberikan banyak manfaat bagi kesehatan. Sebagai contoh, bubur pisang hijau dapat mengobati masalah pencernaan seperti diare dan sembelit pada anak-anak (Falcomer dkk., 2019). Pisang termasuk jenis buah klimaterik yang cepat matang setelah dipanen (Murmu & Mishra, 2018). Gas etilen yang dihasilkan

pada pisang dapat merangsang pematangan yang disertai perubahan warna kulit dari hijau menjadi kuning (Murmu & Mishra, 2018). Terdapat tujuh tahapan perubahan warna kulit pisang seperti yang tertera pada Tabel 2 (Von Loesecke, 1950).

Metode konvensional dalam menentukan kematangan pisang menggunakan indera penglihatan manusia berdasarkan perubahan warna kulit pisang. Namun, metode ini masih memiliki beberapa kekurangan, di antaranya memakan waktu yang lama, bersifat subjektif, dan dapat menghasilkan penilaian yang berbeda bagi setiap individu (Hadfi & Mohd Yusoh, 2018). Oleh karena itu, diperlukan metode yang lebih efisien untuk mencapai pengenalan kematangan pisang yang cepat, konsisten, dan ekonomis. Teknologi *computer vision* dapat menjadi solusi untuk mengklasifikasikan tingkat kematangan buah pisang secara otomatis. Teknologi *computer vision* sering digunakan dalam bidang pertanian (Rico-Fernández dkk., 2019). Pertanian masa depan bertujuan untuk meningkatkan produktivitas, kualitas makanan, dan mengurangi biaya operasional (Fracarolli dkk., 2020). Convolutional Neural Network (CNN) seperti VGG16 (Mishra dkk., 2022), MobileNet V2 (Saragih & Emanuel, 2021), dan CNN yang dikembangkan (Mohamedon dkk., 2021; Saranya dkk., 2022) telah menunjukkan performa yang baik dalam klasifikasi kematangan pisang.

Penelitian sebelumnya telah berhasil menerapkan metode Vision Transformer (ViT) untuk berbagai tugas, seperti: klasifikasi buah strawberry (Zheng dkk., 2022), klasifikasi fraktur femur (Tanzi dkk., 2022), dan klasifikasi Gambar USG Payudara (Gheflati & Rivaz, 2022). Dalam tugas-tugas tersebut, ViT telah menunjukkan keunggulan dalam klasifikasi citra dengan tingkat keakuratan yang tinggi dan kemampuan pemrosesan global yang efisien. Namun, walaupun ViT telah menunjukkan keunggulannya dalam tugas-tugas sebelumnya, keunggulan spesifiknya dalam mengklasifikasikan citra pisang berdasarkan tingkat kematangan belum dijelaskan dengan jelas. Oleh karena itu, penelitian ini bertujuan untuk membandingkan performa model ViT dan Convolutional Neural Network (CNN) dalam mengklasifikasikan tingkat kematangan pisang. Keberhasilan arsitektur *Transformer* pada Natural Language Processing (NLP) menginspirasi model ViT dengan menerapkan arsitektur *Transformer encoder* pada klasifikasi citra (Dosovitskiy dkk., 2020). Eksperimen dilakukan dengan menggunakan lima model ViT yang sudah dilatih atau *pre-trained*, di antaranya: ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32, dan ViT-H/14. Kemudian, *fine-tuning* dilakukan pada dataset tingkat kematangan pisang. Tingkat kematangan pisang dibagi menjadi empat kategori, yaitu mentah, setengah matang, matang, dan terlalu matang, yang masing-masing sesuai dengan tahapan warna kulit 1, 2-4, 5-6, dan 7. Model ini dievaluasi menggunakan *cross-dataset*, di mana pelatihan dan pengujian dilakukan pada dataset yang berbeda. Enam metrik evaluasi dilakukan untuk mengevaluasi performa model, di antaranya: *recall*, *specificity*, *precision*, Negative Predictive Value (NPV), *accuracy*, dan *f1-score*.

2. METODE PENELITIAN

2.1. Vision Transformer (ViT)

Vision Transformer (ViT) adalah model klasifikasi citra yang menggunakan arsitektur *Transformer*. Berdasarkan penelitian (Dosovitskiy dkk., 2020), langkah pertama yang dilakukan ViT adalah membagi citra input menjadi potongan-potongan kecil atau *patch* dengan ukuran yang sama. Untuk citra input $\mathbf{x} \in R^{H \times W \times C}$, dimana H adalah tinggi, W adalah lebar, dan C adalah jumlah *channel* (RGB). Citra input dibagi menjadi potongan-potongan kecil atau *patch* sebanyak $\mathrm{N} = \frac{HW}{P^2}$, dimana P adalah ukuran *patch* (tinggi atau lebar). Kemudian, meratakan atau *flattening patch* menjadi $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$. *Flatten* adalah mengubah matriks C-D menjadi vektor 1-D.

Hasil *flattening* pada *patch* disebut *flattened patches* (x_p). Kemudian, *flattened patches* diubah menjadi D $dimensions$ dengan mengalikannya dengan $embedding\ matrix\ E$ atau *trainable linear projection*. Matriks E ini dibuat secara acak dengan ukuran $((P^2 \cdot C) \times D)$. Hasil perkalian ini disebut *patch embeddings* $(x_p E)$ dengan ukuran $(1 \times D)$. Lalu, untuk membantu bagian klasifikasi, token *learnable* $[class]$ *embedding* $(\mathbf{x}_{class})$ ditambahkan ke *patch embeddings*. Selain itu, untuk menyimpan informasi posisi, *embedding matrix* $(\mathbf{E}_{pos})$ atau *positional embedding* yang dihasilkan secara acak dengan ukuran $((N+1) \times D)$ ditambahkan ke matriks gabungan yang berisi *learnable class embedding* dan *patch embeddings* yang menghasilkan $\mathbf{z}_0$ sesuai pada

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \ldots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{\text{pos}}, \quad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}, \quad \mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

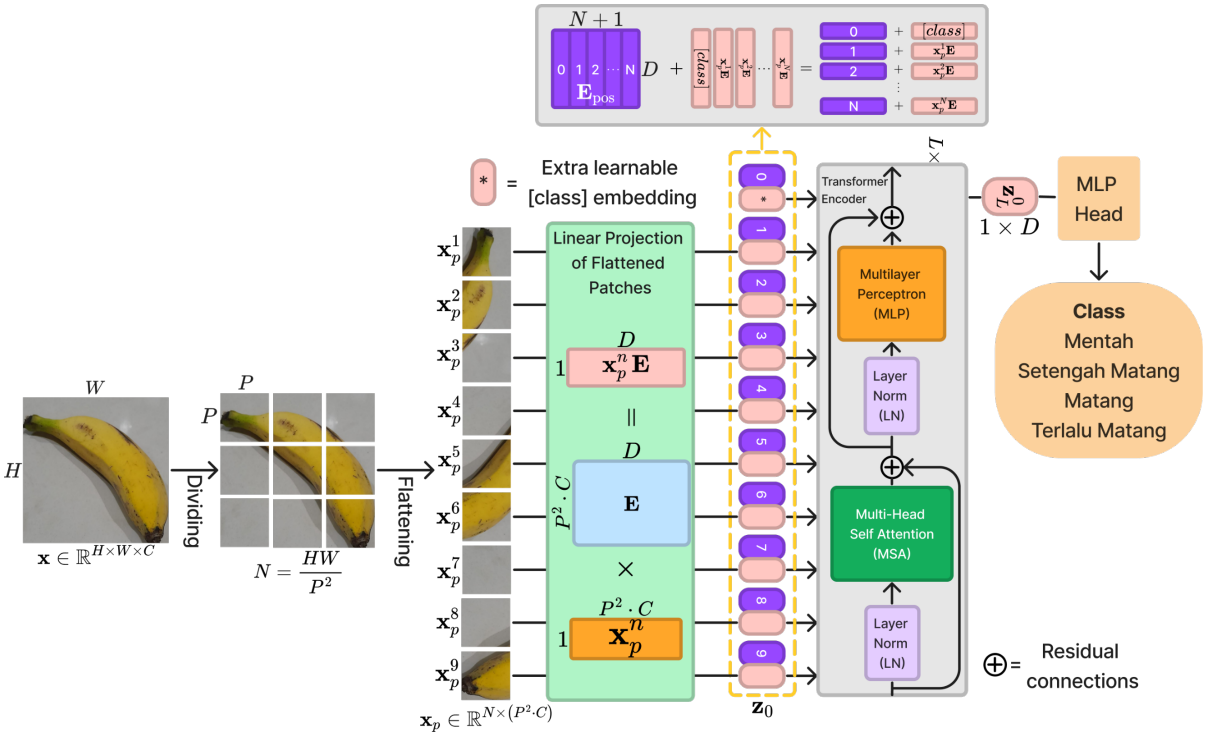
$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \ldots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{\text{pos}}, \quad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}, \quad \mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

$$\mathbf{z}'_\ell = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad \ell = 1 \ldots L \tag{2}$$

$$\mathbf{z}_\ell = \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \quad \ell = 1 \ldots L \tag{3}$$

$$\mathbf{y} = \mathrm{LN}(\mathbf{z}_L^0) \tag{4}$$

Keluaran z_0 dari *embedded patches* diteruskan ke *Transformer encoder*, diulang sebanyak *encoder blocks* (L) di *Transformer encoder*. Komponen *encoder* yaitu Multi-Head Self Attention (MSA) $\mathbf{z}'_\ell = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad \ell = 1 \ldots L$ (2) dan Multilayer Perceptron (MLP) $\mathbf{z}_\ell = \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \quad \ell = 1 \ldots L$ (3), dengan *layer normalization* (LN) dan *residual connections*. Keluaran *Transformer encoder* yang diambil hanya token pertama yaitu



Gambar 1. Arsitektur Vision Transformer

token $[class]$ (z_L^0) dengan ukuran $(1 \times D)$ yang digunakan di MLP head untuk mendapatkan probabilitas masing-masing kategori (y) $\mathbf{y} = \mathrm{LN}(\mathbf{z}_L^0)$ (4). Ilustrasi arsitektur ViT ditunjukan pada Gambar 1.

2.2. *Fine-tune*

Fine-tune adalah proses berdasarkan konsep *transfer learning* dimana model yang dilatih pada tugas tertentu diadaptasi dengan memperbarui semua parameter model untuk tugas baru, sehingga pada dasarnya model dilatih ulang secara keseluruhan. Pemanfaatan model terlatih dalam proses *fine-tuning* dapat mengurangi biaya komputasi, jejak karbon, dan waktu yang diperlukan untuk melatih model tanpa harus melatihnya dari awal (Zhang dkk., 2022).

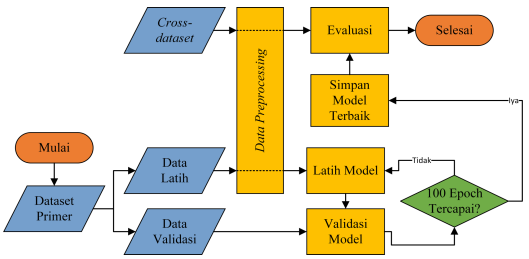
2.3. *Cross-dataset Evaluation*

Cross-dataset evaluation adalah sebuah metode evaluasi performa model *machine learning* atau *deep learning* pada dataset yang berbeda dari dataset latih. Tujuannya adalah untuk mengukur kemampuan model dalam mengeneralisasi hasil prediksi pada data baru yang belum pernah dilihat sebelumnya. Dalam *cross-dataset evaluation*, model dievaluasi dengan menggunakan dataset uji yang berbeda dari dataset latih, sehingga dapat memberikan gambaran tentang sejauh mana model dapat digunakan secara luas dan performa model pada data yang berbeda-beda.

3. PERANCANGAN SISTEM

3.1. Desain Sistem

Dalam penelitian ini, desain sistem dibangun untuk membandingkan model Vision Transformer dan Convolutional Neural Network sebagai model pembanding. Diagram alir dari desain sistem yang dibangun ditunjukkan pada Gambar 2. Sistem dimulai dengan dataset primer yang berukuran 224 × 224 piksel, yang kemudian dibagi menjadi data latih sebesar 80% dan data validasi sebesar 20%. Setelah itu, pra-pemrosesan citra dilakukan dengan menggunakan teknik augmentasi data pada data latih untuk meningkatkan keragaman citra. Teknik augmentasi yang digunakan meliputi rotasi, distorsi perspektif, dan membalik citra. Hal ini bertujuan untuk memperkenalkan variasi pada citra latih dan membantu model dalam beradaptasi dengan perbedaan dalam orientasi, sudut pandang, dan bentuk objek. Data latih hasil augmentasi digunakan untuk melatih kembali model *pre-trained*. Data validasi akan digunakan untuk memvalidasi *hyperparameter* model selama proses pelatihan



Gambar 2. Diagram alir sistem *training* dan evaluasi model

Tabel 1. Detail varian model Vision Transformer

Model	*Patch Size* (*P*)	*Layers* (*L*)	*Hidden Size* (*D*)	MLP *Size*	Heads
ViT-B/16	16	12	768	3072	12
ViT-B/32	32	12	768	3072	12
ViT-L/16	16	24	1024	4096	16
ViT-L/32	32	24	1024	4096	16
ViT-H/14	14	32	1280	5120	16

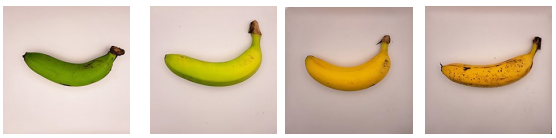
model dilakukan selama 100 *epoch*. Pemilihan 100 *epoch* sebagai jumlah iterasi dalam pelatihan dilakukan untuk memberikan kesempatan model untuk mengenali pola dan fitur yang lebih kompleks pada dataset. Kemudian, model dengan akurasi tertinggi dipilih untuk evaluasi. Selanjutnya, pada tahap pra-pemrosesan citra *cross-dataset* dilakukan dengan mengubah citra menjadi ukuran 224 × 224 piksel, guna mengevaluasi model yang telah dipilih.

3.2. Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari dataset primer dan *cross-dataset* untuk evaluasi model. Dataset primer merupakan data yang diambil langsung oleh peneliti dari sumber aslinya, terdiri dari tingkat kematangan pisang jenis Cavendish dan Ambon. Pisang Cavendish berkontribusi pada tingkat kematangan setengah matang, matang, dan terlalu matang. Namun, untuk tingkat kematangan mentah, sulit ditemukan Pisang Cavendish dengan warna yang sama dengan indeks 1 pada Tabel 2. Oleh karena itu, Pisang Ambon digunakan sebagai pengganti untuk tingkat kematangan mentah dan juga untuk menambahkan warna yang sesuai pada tingkat kematangan terlalu matang. Dataset primer terdiri dari 491 citra mentah, 606 citra setengah matang, 294 citra matang, dan 474 citra terlalu matang, masing-masing sesuai dengan indeks 1, 2-4, 5-6, dan 7 pada Tabel 2. Contoh citra pisang pada dataset primer dapat dilihat pada Gambar 3 dan semua citra pada dataset primer memiliki ukuran (resolusi) 224 × 224 piksel. Dataset primer yang totalnya terdiri dari 1.865 citra dibagi menjadi dua bagian, yaitu 1.490 citra data latih dan 375 citra data validasi, masing-masing sesuai dengan 80% dan 20% dari total dataset.

Tabel 2. Indeks perubahan warna kulit pisang

Indeks	Warna	Tingkat Kematangan
1	Hijau semua	Mentah
2	Hijau dengan sedikit kuning	Setengah Matang
3	Hijau lebih dominan dari kuning	
4	Kuning lebih dominan daripada hijau	
5	Kuning dengan ujung hijau	Matang
6	Kuning penuh	
7	Kuning, berbintik-bintik cokelat	Terlalu Matang



Mentah | Setengah Matang | Matang | Terlalu Matang

Gambar 3. Contoh citra pisang pada dataset primer

Sementara itu, *cross-dataset* digunakan untuk evaluasi model yang dibangun. Dataset ini terdiri dari 5.068 citra pisang hasil penggabungan dari dataset yang dipublikasikan secara *online* oleh (Mazen & Nashat, 2019), (Luciano dkk., 2023), (images.cv, 2022), (gbc, 2023), (Eloise, 2023), dan (Tri Judi Mulajati, 2022). Dataset ini terdiri dari 1.153 citra mentah, 326 citra setengah matang, 1.781 citra matang, dan 1.808 citra terlalu matang. Contoh citra pisang pada *cross-dataset* dapat dilihat pada Gambar 4.

Mentah | Setengah Matang | Matang | Terlalu Matang

Gambar 4. Contoh citra pisang pada *cross-dataset*

3.3. Model

Dalam penelitian ini, digunakan lima variasi model Vision Transformer (ViT), yaitu: ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32, dan ViTH/14. Tabel 1 membandingkan konfigurasi kelima model ViT berdasarkan penelitian (Dosovitskiy dkk., 2020). Model ViT yang digunakan telah dilatih atau *pre-trained* pada dataset ImageNet-21k (14 juta citra, 21.843 kategori) dari Huggingface dan ImageNet-1K (1 juta citra, 1.000 kategori) dari PyTorch. Selain itu, performa model ViT dibandingkan dengan model AlexNet (Krizhevsky, 2014), DenseNet (Huang dkk., 2016), MobileNet V2 (Sandler dkk., 2018), MobileNet V3 (Howard dkk., 2019), ResNet (He dkk., 2015), SqueezeNet (Iandola dkk., 2016), dan VGG (Simonyan & Zisserman, 2015), yang telah dilatih pada dataset ImageNet-1K dari PyTorch.

Selama proses pelatihan, beberapa *hyperparameter* digunakan untuk mengoptimalkan performa model *deep learning*. *Optimizer* yang digunakan adalah AdamW dengan *learning rate* sebesar 2e-05, ukuran *batch* sebesar 8, dan jumlah *epoch* sebanyak 100. Proses *fine-tune* model dilakukan menggunakan GPU A100 dengan memori 40GB yang tersedia di Google Colab.

B adalah *Base*, L adalah *Large*, dan H adalah *Huge*. *Patch size* adalah ukuran *patch* atau citra potongan kecil. *Layers* menunjukkan jumlah *encoder blocks* (*L*) di *Transformer encoder*. *Hidden size* adalah ukuran *embedding D dimension*. MLP *size* adalah jumlah *hidden units* di MLP *layers* pada *Transformer encoder*. Head adalah jumlah *head* di Multi-Headed Self-Attention (MSA) pada *Transformer encoder*.

3.4. Evaluasi

Evaluasi dalam penelitian ini mengevaluasi performa model dengan enam metrik evaluasi, di antaranya: *recall* , *specificity*, *precision*, Negative Predictive Value (NPV), *accuracy*, dan *f1-score.* $Recall = \frac{TP}{TP+FN}$(5) untuk *recall* menghitung persentase dari data positif yang berhasil terdeteksi oleh model, sedangkan $Specificity = \frac{TN}{TN+FP}$ (6) untuk *specificity* mengukur persentase dari data negatif yang berhasil terdeteksi oleh model. $Precision= \frac{TP}{TP+FP}$ (7) untuk *precision* mengukur persentase prediksi positif yang benar dari seluruh prediksi positif yang dilakukan oleh model, sementara $NPV = \frac{TN}{TN+FN}$ (8) untuk NPV mengukur persentase prediksi negatif yang benar dari seluruh prediksi negatif yang dilakukan oleh model. $F_1 = \frac{2*TP}{2*TP+FP+FN}$ (9) untuk *f1-score* merupakan harmonic mean antara recall dan precision, menggabungkan informasi dari kedua metrik tersebut. Terakhir, $OverallAccuracy = \frac{correctlyClassified}{total}$(10) untuk *overall accuracy* menghitung persentase dari total prediksi yang benar dari seluruh hasil prediksi model.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$Specificity = \frac{TN}{TN+FP} \quad (6)$$

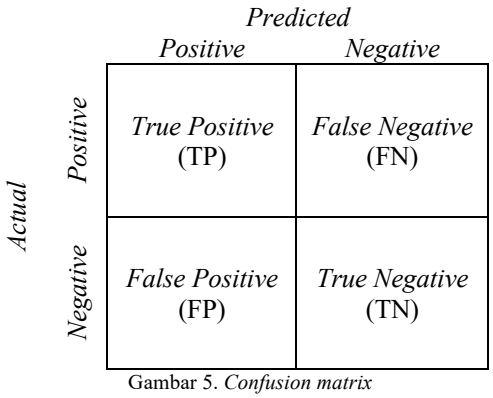
$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$NPV = \frac{TN}{TN+FN} \quad (8)$$

$$F_1 = \frac{2*TP}{2*TP+FP+FN} \quad (9)$$

$$OverallAccuracy = \frac{correctlyClassified}{total} \quad (10)$$

Penelitian ini menggunakan rata-rata ”*macro*” pada persamaan $Recall = \frac{TP}{TP+FN}$ (5),



Gambar 5. *Confusion matrix*

$Specificity = \frac{TN}{TN+FP}$ (6), $Precision= \frac{TP}{TP+FP}$ (7), $NPV = \frac{TN}{TN+FN}$ (8), dan $F_1 = \frac{2*TP}{2*TP+FP+FN}$ (9). Rata-rata ”*macro*” adalah menghitung nilai metrik untuk setiap kelas secara terpisah dan kemudian mengambil rata-rata dari nilai-nilai tersebut. *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) didapatkan dari *Confusion Matrix* yang dapat dilihat pada Gambar 5.

4. HASIL DAN PEMBAHASAN

Bab ini membahas hasil pengujian berdasarkan beberapa skenario sebelumnya, pengujian tersebut meliputi pengujian terhadap model Vision Transformer dan pengujian model Vision Transformer versus model lain.

4.1. Performa Model Vision Transformer (ViT)

Lima model ViT *pre-trained* pada dataset ImageNet-21k dari Huggingface dan dataset ImageNet-1K dari PyTorch telah dilatih ulang (*fine-tuned*) pada dataset primer sebanyak 1.865 citra

Tabel 3. Hasil pelatihan model ViT *pre-trained* pada dataset primer

Model	*Epoch*	Akurasi Tertinggi	Waktu Pelatihan (s)	Ukuran Model (MB)
ViT-B/16-in1k	69	0,9973	2.008	327,4
ViT-B/32-in1k	81	1	1.282	333,7
ViT-L/16-in1k	57	1	5.070	1159,12
ViT-L/32-in1k	59	0,9973	2.302	1167,36
ViT-H/14-in1k	76	0,9947	12.626	2406,4
ViT-B/16-in21k	87	0,992	3.994	343
ViT-B/32-in21k	37	0,9947	3.381	350
ViT-L/16-in21k	81	0,9973	10.953	1244,16
ViT-L/32-in21k	27	0,9973	7.921	1254,08
ViT-H/14-in21k	64	0,9867	24.404	2580,48

Tabel 4. Hasil Hasil evaluasi *cross-dataset* pada *fine-tuned* model ViT

Model	Rerata *Recall*	Rerata *Specificity*	Rerata *Precision*	Rerata NPV	Rerata *F1-score*	Akurasi Keseluruhan
ViT-B/16-in1k	0,8256	0,9363	0,7572	0,9308	0,7681	0,806
ViT-B/32-in1k	0,6862	0,8984	0,682	0,8978	0,6186	0,6871
ViT-L/16-in1k	0,7888	0,9248	0,7461	0,922	0,7338	0,7749
ViT-L/32-in1k	0,6942	0,8956	0,7049	0,8933	0,6101	0,6663
ViT-H/14-in1k	0,7705	0,9177	0,7162	0,9183	0,7243	0,7603
ViT-B/16-in21k	0,8711	0,9638	0,8576	0,9633	0,8636	0,8968
ViT-B/32-in21k	0,8703	0,9576	0,8199	0,9549	0,8188	0,8676
ViT-L/16-in21k	0,8905	**0,9721**	0,8631	**0,9712**	0,8713	**0,9161**
ViT-L/32-in21k	0,7712	0,9266	0,7706	0,9318	0,7167	0,7833
ViT-H/14-in21k	**0,8972**	0,969	**0,9046**	0,9707	**0,8995**	0,9159

pisang dan dievaluasi menggunakan *cross-dataset* sebanyak 5.068 citra pisang.

Tabel 3 menunjukkan hasil pelatihan model *pre-trained* pada dataset primer. Model *fine-tuned* ViT menunjukkan hasil yang sangat baik dengan tingkat akurasi yang tinggi diatas 0,98, yang menunjukkan model tersebut sangat baik dalam mengklasifikasikan kematangan pisang. model ViT *pre-trained* pada dataset ImageNet-1k memiliki waktu pelatihan tercepat dibandingkan model ViT *pre-trained* pada dataset ImageNet-21k. Ukuran *patch* pada model ViT juga mempengaruhi waktu pelatihan dan ukuran model, model dengan ukuran *patch* 16 memiliki waktu pelatihan yang lebih lama dan ukuran model yang lebih kecil dibandingkan model dengan ukuran *patch* 32. Namun, *pre-trained* model ViT pada dataset Imagenet-1k memiliki waktu pelatihan yang lebih cepat dan ukuran model lebih kecil dibandingkan model ViT *pre-trained* pada dataset ImageNet-21k.

Tabel 4 menunjukkan hasil evaluasi *cross-dataset* dari sepuluh model ViT *fine-tuned*. Hasil evaluasi *cross-dataset* model ViT menunjukkan bahwa model ViT-H/14-in21k memiliki performa terbaik dengan rerata *recall*, *precision*, dan *f1-score* tertinggi masing-masing sebesar 0,8972, 0,9046, dan 0,8995. Selain itu, model ViT-L/16-in21k juga memberikan performa yang cukup baik dengan akurasi tertinggi sebesar 0,9161. Model ViT-B/16-in1k memiliki performa terbaik pada *pre-trained* model ViT ImageNet-1K yang memiliki metrik evaluasi lebih tinggi. Sementara itu, model ViT-B/32-in1k dan ViT-L/32-in1k menunjukkan hasil performa yang lebih rendah dibandingkan model lainnya.

4.2. Performa Model Convolutional Neural Networks (CNN)

Sepuluh model CNN *pre-trained* pada dataset ImageNet-1K dari PyTorch telah dilatih ulang (*fine-tuned*) pada dataset primer sebanyak 1.865 citra pisang dan dievaluasi menggunakan *cross-dataset* sebanyak 5.068 citra pisang. Model CNN yang digunakan, di antaranya: AlexNet (Krizhevsky, 2014), DenseNet (Huang dkk., 2016), MobileNet V2 (Sandler dkk., 2018), MobileNet V3 (Howard dkk., 2019), ResNet (He dkk., 2015), SqueezeNet (Iandola dkk., 2016), dan VGG (Simonyan & Zisserman, 2015).

Tabel 5 menunjukkan hasil pelatihan model *pre-trained* pada dataset primer. Semua model CNN menunjukkan tingkat akurasi yang tinggi, dengan

Tabel 5. Hasil pelatihan model CNN *pre-trained* pada dataset primer

Model	*Epoch*	Akurasi Tertinggi	Waktu Pelatihan (s)	Ukuran Model (MB)
Alexnet-in1k	50	0,9893	732	217,5
Densenet121-in1k	100	0,9973	1.795	27,2
Densenet201-in1k	37	0,9947	2.796	70,4
MobileNet V2-in1k	99	0,9947	1.027	8,8
MobileNet V3-in1k	94	0,9333	1.096	16,3
Resnet152-in1k	26	0,9973	1.915	90
Resnet50-in1k	11	0,9973	1.047	222,8
Squeezenet-in1k	97	0,9893	754	2,8
VGG16-in1k	29	0,9947	1.173	512,3
VGG19-in1k	99	0,9973	1.236	532,6

Tabel 6. Hasil evaluasi *cross-dataset* pada *fine-tuned* model CNN

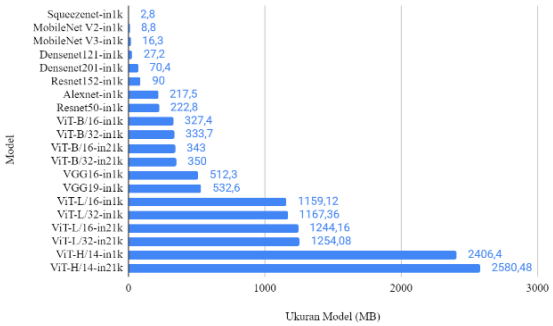
Model	Rerata *Recall*	Rerata *Specificity*	Rerata *Precision*	Rerata NPV	Rerata *F1-score*	Akurasi Keseluruhan
Alexnet-in1k	**0,6888**	0,9166	0,6693	0,9149	**0,6606**	**0,7449**
Densenet121-in1k	0,5467	0,8543	0,5736	0,8642	0,4493	0,5476
Densenet201-in1k	0,6762	0,8972	0,7135	0,8957	0,5935	0,6646
MobileNet V2-in1k	0,4839	0,8166	0,5774	0,8526	0,389	0,4937
MobileNet V3-in1k	0,5221	0,851	0,5215	0,8673	0,4167	0,5383
Resnet152-in1k	0,5172	0,8438	0,5942	0,8717	0,4448	0,5566
Resnet50-in1k	0,5922	0,8649	0,6624	0,8846	0,5093	0,6058
Squeezenet-in1k	0,6872	**0,9196**	**0,7169**	**0,9172**	0,6538	0,7447
VGG16-in1k	0,5834	0,881	0,5579	0,8795	0,5084	0,6186
VGG19-in1k	0,5253	0,838	0,5558	0,8631	0,4249	0,5168

nilai rata-rata di atas 0,98, menunjukkan bahwa model sangat efektif dalam mengklasifikasikan tingkat kematangan pisang. Model Squeezenet-in1k dan MobileNet V2-in1k memiliki ukuran model yang relatif kecil, masing-masing sebesar 2,8 MB dan 8,8 MB. Di sisi lain, model VGG16-in1k dan VGG19-in1k memiliki ukuran model yang jauh lebih besar, masing-masing mencapai 512,3 MB dan 532,6 MB. Waktu pelatihan juga bervariasi; model Alexnet-in1k dan Squeezenet-in1k memiliki waktu pelatihan yang relatif singkat, sekitar 732 detik dan 754 detik. Sedangkan, model Densenet201-in1k memiliki waktu pelatihan paling lama, mencapai 2.796 detik.

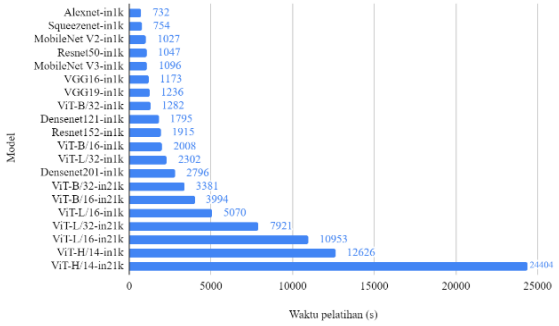
Tabel 6 menunjukkan hasil evaluasi *cross-dataset* dari sepuluh *fine-tuned* model CNN. Hasil evaluasi *cross-dataset* model CNN menunjukkan bahwa model Alexnet-in1k memiliki rerata *recall*, *f1-score*, dan akurasi keseluruhan tertinggi, masing-masing sebesar 0,6888, 0,6606, dan 0,7449. Model Squeezenet-in1k memiliki rerata *specificity*, *precision*, dan NPV tertinggi, masing-masing sebesar 0,9196, 0,7169, dan 0,9172. Sementara itu, model MobileNet V2-in1k dan VGG19-in1k menunjukkan hasil performa yang lebih rendah daripada yang lain.

4.3. Pembahasan

Pada subbab ini, dibahas kinerja model Vision Transformer (ViT) yang dibandingkan dengan model Convolutional Neural Network (CNN) dalam mengklasifikasikan tingkat kematangan pisang. Analisis dilakukan berdasarkan hasil pelatihan dan evaluasi dari kedua jenis model tersebut. Hasil pelatihan dataset primer menunjukkan bahwa baik ViT maupun CNN mampu mencapai tingkat akurasi yang sangat baik. Performa terbaik dari ViT terlihat pada model ViT-B/32-in1k dan ViT-L/32-in1k dengan akurasi 100%, sedangkan model CNN terbaik adalah Densenet121-in1k, Resnet152-in1k, Resnet50-in1k, dan VGG19-in1k dengan akurasi 99,73%. Hal ini menandakan bahwa kedua jenis model tersebut memiliki kemampuan yang kuat dalam mengklasifikasikan kematangan pisang.



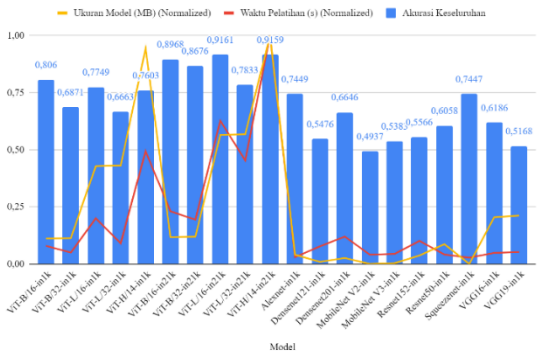
Gambar 6. Ukuran model ViT dan CNN



Gambar 7. Waktu pelatihan ViT dan CNN

Namun, perbedaan ukuran model antara ViT dan CNN cukup signifikan. Gambar 6 menunjukkan bahwa ViT cenderung memiliki ukuran model yang lebih besar daripada CNN. Model *pre-trained* pada ImageNet-1k, ViT-H/14-in1k memiliki ukuran model terbesar dengan 2406,4 MB, sementara model VGG19-in1k adalah model CNN terbesar dengan ukuran 532,6 MB. Selain itu, Gambar 7 menunjukkan bahwa model ViT memerlukan waktu pelatihan yang relatif lebih lama dibandingkan dengan CNN. Model ViT-H/14-in21k membutuhkan waktu pelatihan terpanjang, yakni selama 24.404 detik, sedangkan model Alexnet-in1k memiliki waktu pelatihan tercepat yakni hanya 732 detik.

Gambar 8 menunjukkan hasil evaluasi *cross-dataset* yang memberikan gambaran tentang performa kedua model untuk menilai sejauh mana generalisasi model dan apakah model dapat digunakan secara luas. Berdasarkan hasil evaluasi *cross-dataset fine-tuned* model, terdapat beberapa model ViT yang memiliki akurasi keseluruhan cukup tinggi, yaitu ViT-L/16-in21k dengan akurasi keseluruhan sebesar 0,9161 dan ViT-H/14-in21k dengan akurasi keseluruhan tertinggi kedua sebesar 0,9159. Namun, terdapat juga beberapa model CNN yang memiliki performa cukup baik dalam hal keseluruhan akurasi, seperti AlexNet-in1k dan SqueezeNet-in1k yang masing-masing memiliki akurasi keseluruhan sebesar 0,7449 dan 0,7447.



Gambar 8. Akurasi keseluruhan, waktu pelatihan, dan ukuran model ViT dan CNN

Secara keseluruhan, kedua jenis model, ViT dan CNN, menunjukkan kinerja yang baik dalam

mengklasifikasikan tingkat kematangan pisang. ViT memiliki keunggulan dalam kemampuan generalisasi dan dapat digunakan secara luas. Di sisi lain, CNN memiliki keunggulan dalam ukuran model yang lebih kecil dan waktu pelatihan yang lebih singkat.

5. KESIMPULAN

Dalam penelitian ini, tingkat kematangan pisang dibagi menjadi empat kategori yang masing-masing sesuai dengan tahapan kematangan warna kulitnya. Evaluasi *cross-dataset fine-tuned* model memberikan gambaran tentang performa model dalam menilai sejauh mana generalisasi model dan kemampuan penggunaannya secara luas. Hasil evaluasi pada *cross-dataset* menunjukkan bahwa model ViT-L/16-in21k *pre-trained* pada dataset ImageNet-21k memiliki akurasi keseluruhan tertinggi di antara model Vision Transformer (ViT) yaitu 0,9161 dan model AlexNet *pre-trained* pada dataset ImageNet-1k memiliki akurasi tertinggi di antara model Convolutional Neural Network (CNN) yaitu sebesar 0,7449. Hal ini menunjukkan bahwa model Vision Transformer (ViT) memiliki keunggulan dalam kemampuan generalisasi dan penggunaan model yang luas, sementara CNN memiliki keunggulan dalam ukuran model yang lebih kecil dan waktu pelatihan yang lebih singkat. Aplikasi web dari penelitian ini tersedia di platform HuggingFace Space dengan tautan https://aryap2-klasifikasi-kematangan-pisang.hf.space/. Penelitian selanjutnya disarankan untuk mempertimbangkan beberapa hal yang dapat meningkatkan kualitas hasil yaitu variasi dataset latih ditambahkan dengan menggunakan berbagai sudut pandang dan pencahayaan yang berbeda agar model dapat mempelajari gambar pisang dari berbagai kondisi.

DAFTAR PUSTAKA

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, *abs/2010.11929*. https://arxiv.org/abs/2010.11929

Eloise. (2023). Banana Ripeness Detection Dataset. Dalam *Roboflow Universe*. Roboflow. https://universe.roboflow.com/eloise-pextp/banana-ripeness-detection-o3uia

Falcomer, A. L., Riquette, R. F. R., De Lima, B. R., Ginani, V. C., & Zandonadi, R. P. (2019). Health Benefits of Green Banana Consumption: A Systematic Review. *Nutrients*, *11*(6), 1222. https://doi.org/10.3390/nu11061222

Fracarolli, J. A., Adimari Pavarin, F. F., Castro, W., & Blasco, J. (2020). Computer vision applied to food and agricultural products. *Revista Ciencia Agronomica*, *51*(5), 1–20. https://doi.org/10.5935/1806-6690.20200087

gbc. (2023). Banana ripeness Dataset. Dalam *Roboflow Universe*. Roboflow. https://universe.roboflow.com/gbc-wronj/banana-ripeness-acmx1

Gheflati, B., & Rivaz, H. (2022). Vision Transformers for Classification of Breast Ultrasound Images. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 480–483. https://doi.org/10.1109/EMBC48229.2022.9871809

Hadfi, I. H., & Mohd Yusoh, Z. I. (2018). Banana Ripeness Detection and Servings Recommendation System using Artificial Intelligence Techniques. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(2–8), 83–87. https://jtec.utem.edu.my/jtec/article/view/4464

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *CoRR*, *abs/1512.03385*. http://arxiv.org/abs/1512.03385

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V, & Adam, H. (2019). Searching for MobileNetV3. *CoRR*, *abs/1905.02244*. http://arxiv.org/abs/1905.02244

Huang, G., Liu, Z., & Weinberger, K. Q. (2016). Densely Connected Convolutional Networks. *CoRR*, *abs/1608.06993*. http://arxiv.org/abs/1608.06993

Iandola, F. N., Moskewicz, M. W., Ashraf, K., Han, S., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size. *CoRR*, *abs/1602.07360*. http://arxiv.org/abs/1602.07360

images.cv, I. (2022). Download Banana labeled image classifcation dataset labeled image dataset. Dalam *images.cv*. images.cv. https://images.cv/dataset/banana-image-classification-dataset

Indonesia, S. (2023). *Statistical Yearbook of Indonesia 2023* (D. of Statistical Dissemination, Ed.). BPS-Statistics Indonesia. https://www.bps.go.id/publication/2023/02/28/18018f9896f09f03580a614b/statistik-indonesia-2023.html

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. *CoRR*, *abs/1404.5997*. http://arxiv.org/abs/1404.5997

Luciano, N., de Freitas, E. D. G., Xavier, M. V., Gomes, D. G., & Neves, J. P. H. (2023). *Banana ripeness dataset*. Kaggle. https://www.kaggle.com/dsv/5791191

Mazen, F. M. A., & Nashat, A. A. (2019). Ripeness Classification of Bananas Using an Artificial Neural Network. *Arabian Journal for Science and Engineering*, *44*(8), 6901–6910. https://doi.org/10.1007/s13369-018-03695-5

Mishra, R., Goyal, S., Choudhury, T., & Sarkar, T. (2022). Banana ripeness classification using transfer learning techniques. *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, 1–6. https://doi.org/10.1109/IC3SIS54991.2022.9885244

Mohamedon, M. F., Rahman, F. A., Mohamad, S. Y., & Khalifa, O. O. (2021). Banana Ripeness Classification Using Computer Vision-based Mobile Application. *2021 8th International Conference on Computer and Communication Engineering (ICCCE)*, 335–338. https://doi.org/10.1109/ICCCE50029.2021.9467225

Murmu, S. B., & Mishra, H. N. (2018). Post-harvest shelf-life of banana and guava: Mechanisms of common degradation problems and emerging counteracting strategies. *Innovative Food Science & Emerging Technologies*, *49*, 20–30. https://www.sciencedirect.com/science/article/pii/S146685641730454X

Rico-Fernández, M. P., Rios-Cabrera, R., Castelán, M., Guerrero-Reyes, H. I., & Juarez-Maldonado, A. (2019). A contextualized approach for segmentation of foliage in different crop species. *Computers and Electronics in Agriculture*, *156*, 378–386. https://doi.org/10.1016/j.compag.2018.11.033

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. *CoRR*, *abs/1801.04381*. http://arxiv.org/abs/1801.04381

Saragih, R. E., & Emanuel, A. W. R. (2021). Banana Ripeness Classification Based on Deep Learning using Convolutional Neural Network. *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 85–89. https://doi.org/10.1109/EIConCIT50028.2021.9431928

Saranya, N., Srinivasan, K., & Kumar, S. K. P. (2022). Banana ripeness stage identification: a deep learning approach. *Journal of Ambient Intelligence and Humanized Computing*, *13*(8), 4033–4039. https://doi.org/10.1007/s12652-021-03267-w

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. Dalam Y. Bengio & Y. LeCun (Ed.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1409.1556

Tanzi, L., Audisio, A., Cirrincione, G., Aprato, A., & Vezzetti, E. (2022). Vision Transformer for femur fracture classification. *Injury*, *53*(7), 2625–2634. https://doi.org/https://doi.org/10.1016/j.injury.2022.04.013

Tri Judi Mulajati. (2022). fresh-raw-rotten-banana Dataset. Dalam *Roboflow Universe*. Roboflow. https://universe.roboflow.com/trijudimulajati-student-gunadarma-ac-id/fresh-raw-rotten-banana

Von Loesecke, H. W. (1950). *Bananas: Chemistry, Physiology, Technology*. Interscience Publishers.

Zhang, Y., Zhang, F., & Chen, N. (2022). Migratable urban street scene sensing method based on vision language pre-trained model. *International Journal of Applied Earth Observation and Geoinformation*, *113*, 102989. https://www.sciencedirect.com/science/article/pii/S1569843222001807

Zheng, H., Wang, G., & Li, X. (2022). Identifying strawberry appearance quality by vision transformers and support vector machine. *Journal of Food Process Engineering*, *45*(10), e14132. https://doi.org/https://doi.org/10.1111/jfpe.14132

Halaman ini sengaja dikosongkan