

Peningkatan Akurasi Deteksi Penyakit Malaria Menggunakan *Transfer Learning* pada Arsitektur CNN

DOI: <http://dx.doi.org/10.35889/jutisi.v14i3.3346>

Creative Commons License 4.0 (CC BY – NC)

Sherif Aji Wicaksono^{1*}, Ema Utami²

Teknik Informatika, AMIKOM, Yogyakarta, Indonesia

*e-mail Corresponding Author: adjiesherif@students.amikom.ac.id

Abstract

Malaria is a disease that requires fast and accurate diagnosis. This study compares three CNN architectures VGG16, ResNet50, and InceptionV3 for malaria peripheral blood smear classification using the public NIH dataset. The pipeline includes standardized preprocessing, moderate augmentation, and transfer learning with early stopping (monitoring val_recall). Evaluation on a stratified test set covers accuracy, precision, recall, F1, ROC-AUC, PR-AP, confusion matrix, and paired statistics (McNemar). VGG16 yields the best performance at the 0.50 threshold (AUC 0.9833; AP 0.9846; Recall 0.8955; F1 0.9308) and significantly outperforms InceptionV3 ($\chi^2(1)=111.06$; $p<1\times10^{-24}$). Bootstrap uncertainty (1.000 resamples) gives Recall mean 0.8956 (95% CI 0.8843–0.9077) and F1 mean 0.9309 (95% CI 0.9239–0.9377). Findings support a VGG16-based model as a feasible pre-screening module in resource-constrained settings, emphasizing sensitivity to reduce false negatives.

Keywords: Malaria; Image classification; VGG16; RO-AUC; McNemar

Abstrak

Malaria merupakan penyakit yang memerlukan diagnosis cepat dan akurat. Penelitian ini membandingkan tiga arsitektur CNN VGG16, ResNet50, dan InceptionV3 untuk klasifikasi citra apusan darah tepi malaria berbasis dataset publik NIH. Pipeline meliputi praproses terstandar, augmentasi moderat, dan transfer learning dengan early stopping (monitor val_recall). Evaluasi dilakukan pada himpunan uji terstratifikasi mengukur akurasi, presisi, recall, F1, ROC-AUC, PR-AP, confusion matrix, serta uji statistik berpasangan (McNemar). VGG16 menunjukkan kinerja terbaik pada ambang 0.50 (AUC 0.9833; AP 0.9846; Recall 0.8955; F1 0.9308) dan unggul signifikan atas InceptionV3 ($\chi^2(1)=111.06$; $p<1\times10^{-24}$). Estimasi ketidakpastian berbasis *bootstrap* (1.000 ulangan) menghasilkan Recall mean 0.8956 (CI95% 0.8843–0.9077) dan F1 mean 0.9309 (CI95% 0.9239–0.9377). Temuan ini mendukung model berbasis VGG16 sebagai modul pra-skrining otomatis di lingkungan berdaya komputasi terbatas, dengan penekanan pada sensitivitas untuk meminimalkan salah-negatif.

Kata kunci: Malaria; Klasifikasi citra; VGG16; ROC-AUC; McNemar

1. Pendahuluan

Malaria tetap menjadi isu kesehatan global yang kritis, membutuhkan diagnosis yang cepat, konsisten, dan terukur untuk mencegah komplikasi fatal. Mikroskopi sebagai standar emas diagnosis sangat bergantung pada kualitas preparat, keahlian teknis, ketahanan analis, dan subjektivitas interpretasi, sehingga rentan menghasilkan *false negative* dengan dampak serius [1]. Oleh karena itu, otomatisasi deteksi melalui *deep learning* menjadi solusi potensial untuk meningkatkan akurasi dan kecepatan diagnosis.

Berbagai studi terdahulu telah menunjukkan keberhasilan *Convolutional Neural Network* (CNN) berbasis *transfer learning* dalam klasifikasi citra medis, termasuk deteksi parasit malaria [2][3]. Namun, mayoritas penelitian berfokus pada pelaporan akurasi agregat tanpa menekankan metrik yang krusial dalam konteks skrining medis, seperti *recall* (sensitivitas) untuk meminimalkan kasus terlewat (*false negative*). Selain itu, aspek efisiensi komputasi seperti

ukuran model, waktu pelatihan, dan kebutuhan inferensi yang sangat penting untuk adopsi di fasilitas kesehatan dengan sumber daya terbatas, seringkali kurang mendapat perhatian [4][5].

Berdasarkan *gap* tersebut, penelitian ini mengusulkan evaluasi komparatif yang komprehensif terhadap tiga arsitektur *CNN* populer *VGG16* [6], *ResNet50* [7], dan *InceptionV3* [8] dalam kerangka *transfer learning*. Ketiga arsitektur dipilih karena mewakili filosofi desain yang berbeda seperti *VGG16* dengan konvolusi berulang dan dalam, *ResNet50* dengan koneksi *residual* untuk mitigasi *vanishing gradient*, dan *InceptionV3* dengan modul multi-skala untuk efisiensi komputasi [9][10][11]. Rasionalisasi pemilihan ini adalah untuk mengeksplorasi *trade-off* antara akurasi, sensitivitas (*recall*), dan biaya komputasi dalam konteks yang seragam, sehingga dapat memberikan rekomendasi yang tepat untuk implementasi di lapangan.

Tujuan penelitian ini adalah membangun dan mengevaluasi tiga model klasifikasi biner (terinfeksi vs tidak terinfeksi) untuk deteksi malaria menggunakan *transfer learning* pada arsitektur *VGG16*, *ResNet50*, dan *InceptionV3*, membandingkan kinerja model dengan penekanan khusus pada metrik *recall* dan analisis statistik berpasangan, mengevaluasi efisiensi komputasi masing-masing model dan memberikan *insight* interpretabilitas melalui visualisasi *Gradient-weighted Class Activation Mapping (Grad-CAM)*. Kontribusi penelitian diharapkan tidak hanya pada ranah akademis berupa *benchmark* terstandarisasi, tetapi juga pada ranah praktis berupa rekomendasi model yang feasible sebagai modul pra-skrining otomatis di lingkungan sumber daya terbatas.

2. Tinjauan Pustaka

Tinjauan pustaka difokuskan pada penelitian-penelitian yang mengaplikasikan *deep learning*, khususnya *transfer learning*, untuk deteksi malaria dari citra apusan darah.

Rajaraman et al. (2018) [12] mengevaluasi beberapa arsitektur *CNN* pretrained (sebagai feature extractor dan fine-tuning) pada dataset citra apusan darah tipis malaria. Studi ini melaporkan peningkatan akurasi klasifikasi dan menyoroti potensi transfer learning. Namun, penelitian tersebut tidak melakukan komparasi mendalam antar arsitektur terkait metrik sensitivitas (*recall*) dan efisiensi komputasi, serta tidak menguji signifikansi statistik perbedaan performa model.

Sinha & Gupta (2023) [13] mengimplementasikan transfer learning dengan *ResNet50* untuk diagnosis malaria berbantuan komputer. Mereka melaporkan akurasi yang sangat tinggi (>98%) pada tugas klasifikasi biner. Kelemahan studi ini adalah fokusnya yang terbatas hanya pada satu arsitektur (*ResNet50*) tanpa perbandingan dengan alternatif lain, sehingga sulit menilai keunggulan komparatifnya. Selain itu, analisis terbatas pada akurasi tanpa eksplorasi mendalam terhadap *recall* atau *precision*.

Mujahid et al. (2024) [14] menggunakan arsitektur *EfficientNet* yang lebih modern untuk deteksi malaria pada citra sel darah merah. Mereka mencapai akurasi 97,57% dan menekankan stabilitas model. Meski menggunakan arsitektur efisien, penelitian ini tidak membandingkan kinerjanya secara langsung dengan arsitektur klasik seperti *VGG16* atau *ResNet50* dalam protokol eksperimen yang sama, sehingga *trade-off* antara performa dan efisiensi belum terkuantifikasi secara jelas.

Shahin et al. (2025) [15] mengusulkan kerangka kerja ensemble yang menggabungkan beberapa model (*feature fusion*) untuk deteksi malaria. Pendekatan ini menghasilkan akurasi 96,47% dan *F1-score* 96,45%. Kelemahan utama adalah kompleksitas sistem yang tinggi, baik dalam pelatihan maupun inferensi, yang dapat menjadi kendala untuk penerapan real-time atau di lingkungan dengan sumber daya komputasi terbatas.

Berdasarkan analisis terhadap penelitian terdahulu, penelitian ini berkontribusi pada state of the art dengan beberapa aspek kebaruan pada perbandingan langsung dan adil antara tiga arsitektur *CNN* (*VGG16*, *ResNet50*, *InceptionV3*) pada protokol yang seragam (dataset, pra-proses, augmentasi, skema *fine-tuning*, dan set evaluasi). Hal ini jarang dilakukan secara komprehensif dalam literatur sejenis kemudian bergeser dari fokus tradisional pada akurasi agregat ke penekanan pada *recall* (sensitivitas), metrik yang paling kritis dalam skrining penyakit untuk meminimalkan *false negative*. Analisis juga diperkaya dengan kurva *ROC-AUC*, *PR-AP*, dan *Confusion Matrix* kemudian menggunakan Uji *McNemar* (uji statistik berpasangan non-parametrik) untuk menguji signifikansi perbedaan kinerja model, yang memberikan landasan statistik yang kuat atas klaim keunggulan suatu model. Selain itu, estimasi interval kepercayaan (*CI*) melalui *bootstrap* dilakukan untuk mengkuantifikasi ketidakpastian dari metrik utama (seperti *Recall* dan *F1*), meningkatkan *robustness* pelaporan hasil lalu mengevaluasi

secara eksplisit aspek efisiensi komputasi yang relevan untuk implementasi, seperti ukuran model dalam memori dan estimasi beban komputasi relatif dan menyertakan analisis kualitatif menggunakan *Grad-CAM* untuk memvisualisasikan daerah perhatian model, memberikan *insight* tentang bagaimana model membuat keputusan.

3. Metodologi

3.1 Desain Penelitian

Penelitian bersifat deskriptifkomparatif dan dilaksanakan melalui eksperimen komputasi terkontrol: setiap model dilatih dan diuji pada protokol yang sama agar hasil dapat dibandingkan secara adil. Analisis dilakukan secara kuantitatif pada *testing set* dengan metrik klasifikasi standar. Validitas internal dijaga melalui pemisahan data *train:test* 80:20 yang dilakukan secara acak namun seimbang (*stratified*). Pendekatan transfer learning digunakan untuk melakukan *fine-tuning* model CNN yang telah dilatih awal pada *ImageNet*.

3.2 Dataset dan Sumber Data

Data menggunakan *NIH/LHNCBC Malaria Cell Images* yang luas dipakai sebagai benchmark terbuka. Dataset berisi 27.558 citra mikroskopik sel darah, terbagi merata antara kelas *Parasitized* (13.779) dan *Uninfected* (13.779), dengan label yang jelas. Data diunduh dari repositori resmi *NIH*, diekstrak, dan diperiksa untuk memastikan tidak ada berkas rusak atau tidak relevan. Seluruh citra diproses seragam menjadi 100×100 piksel (*RGB*). Prapemrosesan meliputi: *resize* ke 100×100, normalisasi piksel ke rentang 0.01.0, *label encoding* (*Parasitized*=1; *Uninfected*=0), dan augmentasi (flipping horizontal, rotasi, *zoom*, perubahan kecerahan) yang hanya diterapkan pada *training set*. Pembagian 80% train / 20% test dilakukan *stratified* menggunakan *scikit-learn* dan *ImageDataGenerator*.

3.3. Lingkungan Komputasi

Seluruh proses pengolahan data dan pelatihan model dilakukan pada *Google Colaboratory (Colab)* dengan dukungan *GPU* untuk efisiensi pelatihan. *Library* utama: *TensorFlow* dan *Keras*, didukung *NumPy*, *OpenCV*, dan *matplotlib*.

3.4. Arsitektur dan Penyetelan Model

Tiga arsitektur pretrained (*VGG16*, *ResNet50*, *InceptionV3*) diimpor dari *Keras Applications* tanpa top layer. Pada bagian atas ditambahkan *custom classifier* yang terdiri dari *GlobalAveragePooling2D*, *Dense* beraktivasi *ReLU*, *Dropout* untuk regularisasi, dan *Dense* keluaran beraktivasi *sigmoid* untuk klasifikasi biner. *Optimizer Adam* dengan learning rate 0,0001 dan *loss Binary Crossentropy* digunakan secara konsisten pada semua model.

3.5. Protokol Pelatihan

Replikasi dikendalikan dengan penetapan random seed 42 pada *Python*, *NumPy*, dan *TensorFlow*. *Early stopping* diaktifkan dengan monitor *val_recall*, *patience* 5, dan pemulihan bobot terbaik. Pelatihan dibatasi pada maksimal 30 *epoch*; jumlah epoch aktual dilaporkan pada bagian Hasil. Pengukuran dilakukan pada input 100×100 *RGB*, *batch size* 32, *optimizer Adam* (1e4). Lingkungan komputasi didokumentasikan (versi *TensorFlow/Keras*, *GPU/CPU* yang digunakan) untuk transparansi evaluasi. Selain pelaporan pada ambang 0,50, dilakukan penelusuran ambang untuk memaksimalkan *recall* dengan kendala presisi minimum 0,90, berlandaskan kebutuhan skrining yang memprioritaskan penekanan *false negative*.

3.6. Protokol Pengujian dan Metrik Evaluasi

Evaluasi pada *testing set* dilakukan menggunakan metrik: Akurasi, Presisi, Recall (Sensitivitas), F1-score, serta Confusion Matrix. Hasil tiap model disajikan dalam bentuk tabel dan grafik performa, termasuk visualisasi *confusion matrix*, untuk memudahkan perbandingan.

3.7. Analisis Komparatif dan Statistik

perbandingan kinerja antar-model dilakukan dengan uji McNemar pada luaran biner (benar/salah) tiap sampel. Uji ini menguji hipotesis nol bahwa probabilitas kesalahan kedua model adalah sama, menggunakan tabel berpasangan (*a, b, c, d*) dan statistik χ^2 dengan koreksi kontinuitas. Selain nilai *p*, ukuran efek yang relevan: proporsi pasangan diskordan yang

dimenangkan model A ($\frac{c}{b+c}$) dan matched-pair odds ratio $OR = \frac{c}{b}$ beserta CI/95% (pendekatan log-normal).

4. Hasil dan Pembahasan

4.1. Realisasi Dataset dan Pembagian Data

Dataset yang digunakan berisi 27.558 citra, terdiri atas 13.779 kelas *Parasitized* dan 13.779 kelas *Uninfected*. Dengan skema traintest 80:20 terstratifikasi:

1) Training set: 22.046 citra

Parasitized: 11.023 - *Uninfected*: 11.023

2) Testing set: 5.512 citra

Parasitized: 2.756 - *Uninfected*: 2.756

Angka-angka ini bersifat deterministik dari total data dan proporsi split, sehingga dapat dipakai sebagai acuan tetap untuk seluruh eksperimen komparatif.

4.2. Kapasitas Model dan Memori (FP32)

Perhitungan jumlah parameter dan estimasi ukuran bobot (FP32 $\approx$ 4 byte/parameter) terdapat pada tabel 1.

Tabel 1. Perhitungan Jumlah Parameter

Arsitektur	Parameter ($\approx$)	Perkiraan Ukuran Bobot
VGG16	138.357.544	$\pm$ 527,8 MB
ResNet50	25.636.712	$\pm$ 97,8 MB
InceptionV3	23.851.784	$\pm$ 91,0 MB

Keterbatasan memori/penyimpanan, *ResNet50* dan *InceptionV3* relatif lebih ramah dibanding *VGG16*.

4.3 Estimasi Beban Komputasi Relatif (Berbasis Skala Input)

GFLOPs absolut bergantung detail implementasi, beban komputasi konvolusi skala pertama dapat didekati dengan rasio luas input:

1) Untuk model *native* 224×224 (*VGG16*, *ResNet50*), input $100 \times 100 = 19,93\%$ dari luasan 224×224 .

2) Untuk *native* 299×299 (*InceptionV3*), input $100 \times 100 = 11,19\%$ dari luasan 299×299 .

Pada resolusi 100×100 , *ResNet50* dan *VGG16* mengalami pengurangan beban komputasi hingga = 80%, sementara *InceptionV3* hingga = 89% dari *native resolution*-nya. Ini sejalan dengan tujuan efisiensi komputasi di lingkungan terbatas.

4.4 Realisasi Proses Pelatihan (Langkah per *Epoch*)

Dengan *batch size* = 32 dan $|train| = 22.046$, jumlah langkah per *epoch* (dibulatkan ke atas):

1) Steps per *epoch* = $\text{ceil}(22.046 / 32) = 689$ langkah/epoch

2) Jika durasi pelatihan direncanakan 30 *epoch*, total pembaruan gradien = 20.670 langkah.

Untuk pelaporan efisiensi (waktu per epoch) dan reproduceability (*early stopping* berhenti di epoch ke-e dari 30).

4.5 Hasil Uji

Hasil uji performa pada Testing Set dan Confusion Matrix per Model ($n = 5.512$; positif = 2.756; negatif = 2.756) ditampilkan pada tabel 2 dan tabel 3.

Tabel 2. Performa Testing Set

Model	Akurasi	Presisi (Pos)	Recall (Pos)	F1 (Pos)	AUC	AP
<i>VGG16</i>	0,9334	0,9690	0,8955	0,9308	0,9833	0,9846
<i>ResNet50</i>	0,5372	0,5198	0,9775	0,6787	0,7675	0,8136
<i>InceptionV3</i>	0,8862	0,9134	0,8534	0,8824	0,9546	0,9599

Tabel 3. Confusion Matrix per Model

Model	TP	FP	FN	TN
VGG16	2468	79	288	2677
ResNet50	2694	2489	62	267
InceptionV3	2352	223	404	2533

Perhitungan dengan penerapan persamaan yang digunakan:

VGG16 (TP=2468, FP=79, FN=288, TN=2677)

$$Akurasi = \frac{2468 + 2677}{5512} = \frac{5145}{5512} = 0,9334$$

$$Presisi = \frac{2468}{2468 + 79} = \frac{2468}{2547} = 0,9690$$

$$Recall = \frac{2468}{2468 + 288} = \frac{2468}{2756} = 0,8955$$

$$F1 = \frac{2 \times 2468}{2 \times 2468 + 79 + 288} = \frac{4936}{5303} = 0,9308$$

$$Spesifisitas/TNR = \frac{2677}{2677 + 79} = \frac{2677}{2756} = 0,9713$$

$$FPR = \frac{79}{79 + 2677} = \frac{79}{2756} = 0,0287$$

$$FNR = \frac{288}{2468 + 288} = \frac{288}{2756} = 0,1045$$

$$AUC = 0,9833, AP = 0,9846$$

ResNet50 (TP=2694, FP=2489, FN=62, TN=267)

$$Akurasi = \frac{2694 + 267}{5512} = \frac{2961}{5512} = 0,5372$$

$$Presisi = \frac{2694}{2694 + 2489} = \frac{2694}{5183} = 0,5198$$

$$Recall = \frac{2694}{2694 + 62} = \frac{2694}{2756} = 0,9775$$

$$F1 = \frac{2 \times 2694}{2 \times 2694 + 2489 + 62} = \frac{5388}{7939} = 0,6787$$

$$Spesifisitas/TNR = \frac{267}{267 + 2489} = \frac{267}{2756} = 0,0969$$

$$FPR = \frac{2489}{2489 + 267} = \frac{2489}{2756} = 0,9031$$

$$FNR = \frac{62}{2694 + 62} = \frac{62}{2756} = 0,0225$$

$$AUC = 0,7675, AP = 0,8136$$

InceptionV3 (TP=2352, FP=223, FN=404, TN=2533)

$$Akurasi = \frac{2352 + 2533}{5512} = \frac{4885}{5512} = 0,8862$$

$$Presisi = \frac{2352}{2352 + 223} = \frac{2352}{2575} = 0,9134$$

$$Recall = \frac{2352}{2352 + 404} = \frac{2352}{2756} = 0,8534$$

$$F1 = \frac{2 \times 2352}{2 \times 2352 + 223 + 404} = \frac{4704}{5331} = 0,8824$$

$$Spesifisitas/TNR = \frac{2533}{2533 + 223} = \frac{2533}{2756} = 0,9191$$

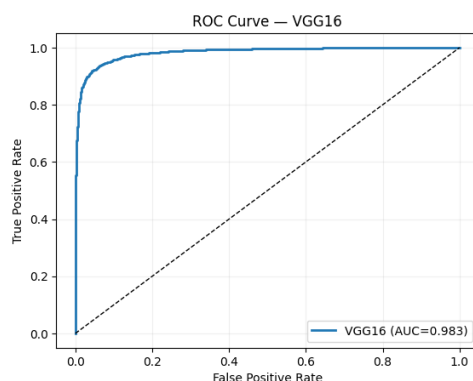
$$FPR = \frac{223}{223 + 2533} = \frac{223}{2756} = 0,0809$$

$$FNR = \frac{404}{2352 + 404} = \frac{404}{2756} = 0,1466$$

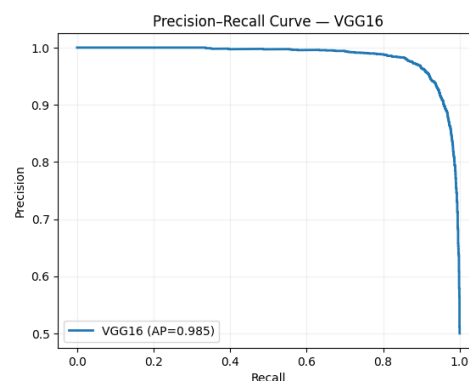
$$AUC = 0,9546, \quad AP = 0,9599$$

Visualisasi ROC dan PR

Digambarkan kurva ROC untuk model terbaik (VGG16) pada *testing set* dan mendapatkan AUC sebesar 0,9833 pada gambar 1. Selain kurva ROC, gambar 2 merupakan kurva Precision-Recall untuk menilai kinerja pada kelas positif. *Reliability diagram* menunjukkan kalibrasi probabilitas yang memadai. Kurva metrik vs ambang (*Presisi*, *Recall*, *F1*, *Spesifisitas*) menginformasikan *trade-off* operasional.



Gambar 1. Kurva ROC VGG16



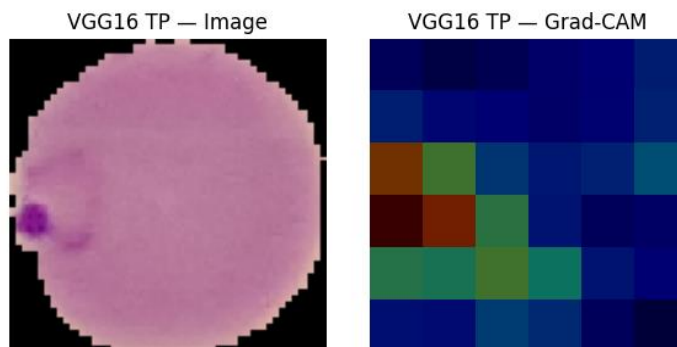
Gambar 2. Kurva PR VGG16

Hasil Bootstrap CI

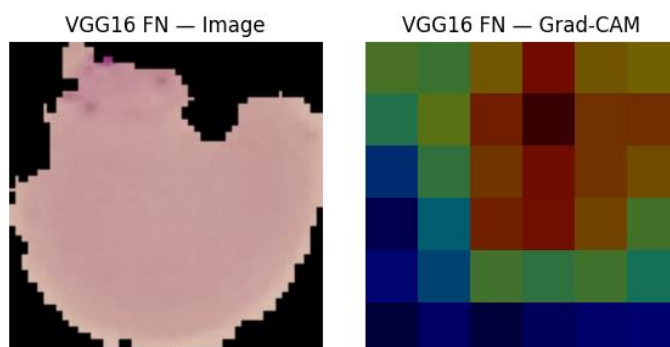
Pada ambang 0,50, VGG16 memperoleh AUC 0,9833, AP 0,9846, Recall 0,8955, dan F1 0,9308. Estimasi ketidakpastian dengan *bootstrap* (1.000 ulangan) menunjukkan Recall mean 0,8956, CI95% [0,8843; 0,9077] dan F1 mean 0,9309, CI95% [0,9239; 0,9377], menandakan stabilitas kinerja pada titik operasi yang dilaporkan.

Visualisasi Grad-CAM

Memvisualisasikan Grad-CAM pada *true positive* (gambar 3) dan *false negative* (gambar 4) untuk memeriksa apakah *attention* model berfokus pada area sitoplasma eritrosit yang relevan dengan keberadaan parasit.



Gambar 3. VGG16 TP Image dan Grad-CAM



Gambar 4. VGG16 FN Image dan Grad-CAM

Grad-CAM

Grad-CAM pada *TP* memperlihatkan pemusatan perhatian pada area eritrosit dengan indikasi parasit; pada *FN*, peta panas cenderung menyebar dan melemah pada wilayah target. Visual ini memperkuat interpretasi bahwa *VGG16* mempertahankan detail lokal yang relevan bagi deteksi parasit, selaras dengan *recall* yang tinggi.

Uji McNemar

Definisi pada sampel uji yang sama di ambang 0,50:

- 1) a: keduanya benar
- 2) b: *VGG16* salah, *InceptionV3* benar
- 3) c: *VGG16* benar, *InceptionV3* salah
- 4) d: keduanya salah

Untuk menghitung a dan d memakai total benar masing-masing model dan total *N*:

- 1) Total benar *VGG16* = $TP+TN = 2468+2677 = 5145 \rightarrow$ ini sama dengan $a + c$
- 2) Total benar *InceptionV3* = $2352+2533 = 4885 \rightarrow$ ini sama dengan $a + b$
- 3) Total sampel uji $N = TP+FP+FN+TN = 5512 \rightarrow$ ini sama dengan $a + b + c + d$
- 4) Diketahui dari nilai ambang 0,50 dihasilkan $b=172$ dan $c=432$

Maka:

$$a = (a + b) - b = 4885 - 172 = 4713$$

$$d = N - (a + b + c) = 5512 - (4713 + 172 + 432) = 195$$

Tabel 4. 2x2 McNemar (Berpasangan)

	InceptionV3 Benar	InceptionV3 Salah
VGG16 Benar	a = 4713	c = 432
VGG16 Salah	b = 172	d = 195

$$a + c = 5145 \text{ (VGG16)}, a + b = 4885 \text{ (Inceptionv3)}, a + b + c + d = 5512$$

Statistik Uji McNemar dengan koreksi kontinuitas

$$X^2 = \frac{(|b - c| - 1)^2}{b + c}$$

$$X^2 = \frac{(|172 - 432| - 1)^2}{172 + 432} = \frac{(260 - 1)^2}{604} = \frac{259^2}{604} = 111,0613$$

$$p - \text{value untuk } X^2 df = 1$$

$$p = \Pr(X_{(1)}^2 \geq 111,0613) = 5,74 \times 10^{-26}$$

$$p < 1 \times 10^{-24} \text{ atau } p < 0,001$$

Ukuran Diskordan

$$\text{Jumlah diskordan} = b + c = 172 + 432 = 604$$

$$\text{Winrate VGG16} = \frac{c}{b + c} = \frac{432}{604} = 0,7152 \text{ (71,5\%)}$$

$$\text{Matched - pair odds ratio (OR)} = OR = \frac{c}{b} = \frac{432}{172} = 2,5116$$

CI95% OR (aproksimasi log - normal)

$$= \ln(OR) \pm 1,96 \sqrt{\frac{1}{b} + \frac{1}{c}} \Rightarrow \text{CI95\%} = [2,105; 2,997]$$

4.6 Pembahasan

Hasil menunjukkan bahwa VGG16 mencapai kinerja terbaik secara keseluruhan pada ambang 0,50, dengan AUC 0,9833, AP 0,9846, Recall 0,8955, dan F1 0,9308. Uji McNemar mengkonfirmasi bahwa perbedaan kinerja antara VGG16 dan InceptionV3 adalah signifikan secara statistik ($\chi^2(1)=111.06$; $p<1 \times 10^{-24}$), dengan VGG16 memenangkan 71,5% dari kasus diskordan (OR=2,51). Estimasi *bootstrap* yang menghasilkan Recall (CI95% 0,8843–0,9077) dan F1 (CI95% 0,9239–0,9377) mengindikasikan stabilitas performa VGG16.

Temuan tingginya recall VGG16 (0,8955) sangat relevan dalam konteks klinis skrining malaria, di mana meminimalkan *false negative* (kasus terinfeksi yang terlewat) adalah prioritas utama [1]. Hasil melengkapi temuan Rajaraman et al. [12] menunjukkan keberhasilan *transfer learning*, penelitian ini memberikan hasil kuantitatif melalui perbandingan langsung dan uji statistik. Sementara Sinha & Gupta [13] melaporkan akurasi >98% dengan ResNet50, penelitian ini menemukan performa ResNet50 dengan (Akurasi 0,5372) karena tingginya *false positive*. Disparitas ini mungkin disebabkan oleh perbedaan protokol *fine-tuning*, augmentasi, atau skema evaluasi.

Dibandingkan dengan pendekatan *ensemble* yang kompleks seperti yang diusulkan Shahin et al. [15] (Akurasi 96,47%, F1 96,45%), VGG16 tunggal dalam penelitian ini mencapai F1 yang sebanding (0,9308) dengan kompleksitas dan kebutuhan komputasi yang lebih rendah. Hal ini menjadikan VGG16 sebagai kandidat praktis untuk implementasi di lingkungan sumber daya terbatas. Meskipun Mujahid et al. [14] menggunakan arsitektur yang lebih efisien (EfficientNet), penelitian ini menunjukkan arsitektur klasik seperti VGG16 masih kompetitif ketika dioptimalkan dengan protokol *fine-tuning* dan evaluasi yang berfokus pada recall.

Visualisasi Grad-CAM memberikan penjelasan mengapa VGG16 berkinerja baik pada *true positive*, model secara konsisten memusatkan pada area *sitoplasma eritrosit* yang mengandung parasit, sedangkan pada *false negative* perhatiannya cenderung menyebar. Namun, terdapat *trade-off* yang perlu dipertimbangkan meskipun unggul dalam recall dan AUC, VGG16 memiliki ukuran model terbesar ($\pm 528 \text{ MB}$) yang mungkin menjadi kendala untuk penyebaran di perangkat *edge* dengan memori terbatas. Untuk skenario tersebut, arsitektur seperti EfficientNet atau MobileNet dengan kalibrasi ulang ambang klasifikasi untuk mempertahankan recall tinggi, dapat menjadi alternatif yang lebih baik.

5. Simpulan

Penelitian ini menunjukkan bahwa VGG16 merupakan model paling andal untuk klasifikasi citra sel darah tepi malaria pada dataset NIH dalam ruang lingkup dan protokol evaluasi yang digunakan. Pada ambang 0,50, VGG16 mencapai *AUC* 0,9833, *AP* 0,9846, *Recall* 0,8955, dan *F1* 0,9308, serta secara statistik unggul atas InceptionV3 berdasarkan uji McNemar $\chi^2(1)=111,0613$; $p<1\times10^{-24}$. Estimasi ketidakpastian melalui *bootstrap* menegaskan stabilitas performa (*Recall mean* 0,8956; *CI*95% 0,8843-0,9077 dan *F1 mean* 0,9309; *CI*95% 0,92390,9377). Secara praktis, kombinasi sensitivitas tinggi dan keseimbangan presisi-recall ini mengindikasikan bahwa VGG16 layak dipertimbangkan sebagai modul pra-skrining otomatis untuk mempercepat alur kerja di fasilitas laboratorium berdaya terbatas.

Ke depan, pekerjaan lanjut yang disarankan mencakup validasi eksternal pada citra lapangan dengan variasi pewarnaan, analisis *robustnes* terhadap pergeseran domain, penalaran ambang berbasis target (*recall-first* dengan batas presisi minimum), serta eksplorasi arsitektur efisien (*MobileNetV3/EfficientNetV2*) untuk menurunkan latensi inferensi pada perangkat terbatas. Dengan langkah-langkah tersebut, penerapan sistem skrining otomatis berbasis VGG16 berpotensi lebih kuat dan terukur di lingkungan penggunaan nyata.

Daftar Referensi

- [1] W. Health Organization, *World malaria report 2023 -- spread view*. London: World Health Organization, 2023.
- [2] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. Thoma, "Image analysis and machine learning for detecting malaria," *Transl. Res.*, vol. 194, pp. 36–55, Apr. 2018.
- [3] M. S. Davidson *et al.*, "Automated detection and staging of malaria parasites from cytological smears using convolutional neural networks.," *Biol. imaging*, vol. 1, p. e2, 2021.
- [4] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- [5] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, vol. 6, no. 1, pp. 1–14, Sep. 2015.
- [6] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, vol. 2, no. 1, pp. 1–22, Oct. 2021.
- [7] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," vol. 1, no. 1, pp. 1–8, Dec. 2017.
- [8] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data* 2019 61, vol. 6, no. 1, pp. 1–48, Jul. 2019.
- [9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2015.
- [10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 2818–2826, Dec. 2015.
- [11] M. Tan and Q. V Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2019, pp. 6105–6114.
- [12] S. Rajaraman *et al.*, "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images Sivaramakrishnan," *PeerJ*, vol. 2018, no. 4. PeerJ Inc., p. e4568, 16-Apr-2018.
- [13] S. Sinha and N. Gupta, "Computer-aided Diagnosis of Malaria through Transfer Learning using the ResNet50 Backbone," *PHARMBIT*, vol. 36, pp. 1–9, Apr. 2023.
- [14] M. Mujahid *et al.*, "Efficient deep learning-based approach for malaria detection using red blood cell smears," *Sci. Rep.*, vol. 14, no. 1, pp. 1–16, Dec. 2024.
- [15] O. R. Shahin, H. H. Alshammari, R. N. Alabdali, A. M. Salaheldin, and N. Saleh, "Automated multi-model framework for malaria detection using deep learning and feature fusion," *Sci. Rep.*, vol. 15, no. 1, pp. 1–18, Dec. 2025.