Klasifikasi Level Retinopati Diabetik Menggunakan Metode *Hybrid Vision Transformer* dan *EfficientNet*

Meakhel Gunawan[1], Yuni Yamasari[2]

[1,2] Program Studi S1 Teknik Informatika, Universitas Negeri Surabaya
[1]meakhel.22116@mhs.unesa.ac.id
[2]yuniyamasari@unesa.ac.id

***Abstrak*—Retinopati diabetik merupakan komplikasi diabetes melitus yang menjadi salah satu penyebab utama kebutaan di dunia, termasuk di Indonesia. Oleh sebab itu, deteksi dan klasifikasi tingkat keparahan secara dini berbasis citra fundus retina menjadi krusial untuk mendukung penanganan klinis yang tepat. *Vision Transformer* yang memiliki keunggulan dalam menangkap konteks global citra, dan *EfficientNet* yang unggul dalam mengekstraksi fitur lokal, memiliki peluang untuk dikombinasikan secara optimal dalam klasifikasi level retinopati diabetik. Menggunakan *Dataset APTOS 2019 Blindness Detection* dengan tahapan *preprocessing* berupa *cropping*, *resize*, CLAHE, dan normalisasi, serta augmentasi data untuk mengatasi ketidakseimbangan kelas, model MobileViT-XS yang dikombinasikandengan *EfficientNet*-B0 hingga *EfficientNet*-B4 diuji dan dievaluasi menggunakan metrik efektivitas serta metrik efisiensi komputasi. Hasil penelitian menunjukkan bahwa *hybrid* MobileViT-XS dan *EfficientNet*-B1 dengan *resize* 512×512 piksel memberikan kinerja paling optimal, dengan akurasi validasi sebesar 91,80% dan akurasi pengujian sebesar 93,24%, serta efisiensi komputasi yang seimbang dengan waktu pelatihan 101 menit 59.44 detik, FLOPs sekitar 6,80G, dan ukuran model 37,95 MB. Penelitian ini menyimpulkan bahwa pendekatan metode *hybrid* mampu menghasilkan model yang efektif dan efisien untuk mendukung sistem deteksi dini retinopati diabetik berbasis kecerdasan buatan.**

Kata Kunci*—retinopati diabetik, *vision transformer*, *EfficientNet*, *hybrid model*, klasifikasi citra, *deep learning

I. PENDAHULUAN

Retinopati Diabetik (RD) adalah komplikasi mikrovaskular yang parah akibat diabetes melitus, yang secara progresif merusak pembuluh darah di retina dan menjadi salah satu penyebab utama kebutaan pada orang dewasa di seluruh dunia [1]. Laporan terbaru dari *International Diabetes Federation* (IDF) mengestimasi sekitar 10,5%, atau sekitar 1 dari 10 orang, populasi manusia dewasa saat ini hidup dengan penyakit Diabetes Melitus. Pada tahun 2021, Indonesia menjadi negara kelima dengan penderita diabetes terbanyak di dunia sejumlah 19,5 juta orang dewasa umur 20 hingga 79 tahun. Selain itu, diproyeksikan bahwa tahun 2045, penderita diabetes di Indonesia naik 46% menjadi 28,6 juta orang [2]. Prevalensi Diabetes Melitus ini juga sejalan dengan penderita RD yang terjadi pada 30% hingga 40% dari seluruh penderita diabetes. Analisis meta terbaru juga memperkirakan bahwa saat ini terdapat sekitar 103 juta orang yang mengalami RD secara global, dan angka ini diperkirakan akan meningkat hingga mencapai 161 juta orang pada tahun 2045 [3]. Oleh sebab itu, deteksi dini dan intervensi yang tepat waktu menjadi sangat krusial untuk menangani masalah mata maupun dalam mencegah komplikasi hingga kehilangan penglihatan permanen akibat RD.

Meskipun *Convolutional Neural Network* (CNN) telah menjadi standar dalam klasifikasi citra medis karena terbukti unggul dalam menangkap dan mengekstraksi fitur lokal, CNN masih memiliki keterbatasan dalam memahami konteks global maupun mengabaikan bagian-bagian yang relevan yang mengakibatkan hilangnya informasi hubungan spasial yang penting untuk mendeteksi tingkat keparahan RD [4][5]. Seiring dengan perkembangan CNN, *Vision Transformer* (ViT) telah muncul sebagai arsitektur yang menjanjikan dalam bidang *computer vision* dengan memperlakukan gambar sebagai rangkaian patch dan menggunakan mekanisme *self-attention* untuk menangkap dependensi jarak jauh dan konteks global dalam citra [6].

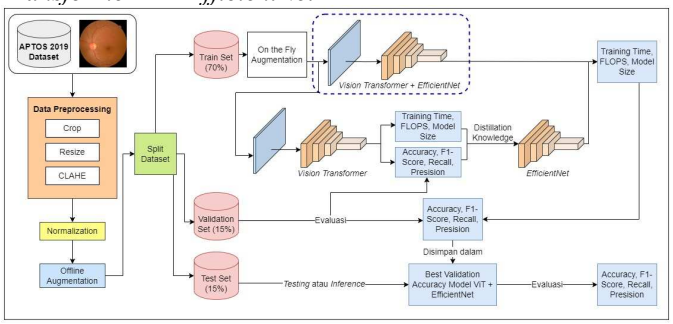
Kondisi tersebut sekaligus membuka peluang untuk mengeksplorasi pendekatan baru yang lebih efisien dengan menggabungkan keunggulan CNN melalui ekstraksi fitur lokal dengan kemampuan ViT dalam menangkap konteks global. Studi yang dilakukan oleh Rautaray et al. [7] mengusulkan kerangka kerja *knowledge distillation* dengan FastViT-MA36 sebagai *teacher* dan *EfficientNet*-B0 sebagai *student* untuk klasifikasi tingkat keparahan RD, yang mencapai akurasi 95.39% dengan biaya komputasi 0.38 GFLOPs dan 42.7 juta parameter. Penelitian lain oleh Fu et al. [8], mengembangkan MSEF-Net dengan menggabungkan fitur *multi-scaling* dari *EfficientNet* dan *attention module*, berhasil meraih akurasi 97.5% pada dataset Messidor1. Sementara itu, studi oleh Tanwar et al. [9], menunjukkan bahwa penggabungan *EfficientNet*-B0 dan ViT mampu mencapai akurasi hingga 99.82% dengan jumlah parameter sekitar 5.1 juta untuk klasifikasi penyakit gastrointestinal. Ketiga penelitian ini menegaskan bahwa pendekatan berbasis kombinasi CNN–ViT, baik melalui *hybrid* maupun *knowledge distillation*, memiliki potensi besar untuk menghasilkan sistem klasifikasi medis yang akurat sekaligus relevan bagi penerapan nyata pada perangkat dengan keterbatasan sumber daya.

Oleh sebab itu, penelitian ini mengusulkan pendekatan model *hybrid* yang menggabungkan keunggulan CNN, yang diwakilkan oleh *EfficientNet*, dalam mengekstrak fitur lokal dengan kemampuan *Vision Transformer*, yang diwakilkan oleh MobileViT-XS, dalam menangkap konteks global. Melalui metode ini, diharapkan tercipta keseimbangan yang optimal antara akurasi tinggi dan efisiensi komputasi, serta mampu memberikan kinerja unggul dengan jumlah parameter dan kebutuhan komputasi (FLOPs) yang relatif kecil. Selain itu, diharapkan juga hasil penelitian ini tidak hanya menambah wawasan dalam bidang akademis, melainkan juga menjawab

kesenjangan riset terkait efektivitas dan efisiensi *hybrid Vision Transformer* dengan *EfficientNet* dalam klasifikasi level Retinopati Diabetik.

II. METODOLOGI PENELITIAN

Pada Gambar 1 membeirkan gambaran mengenai kerangka alur kerja secara keseluruhan dari metode *hybrid Vision Transformer* dan *EfficientNet*.



Gbr. 1 Diagram Alur Penelitian.

A. Perangkat yang Digunakan

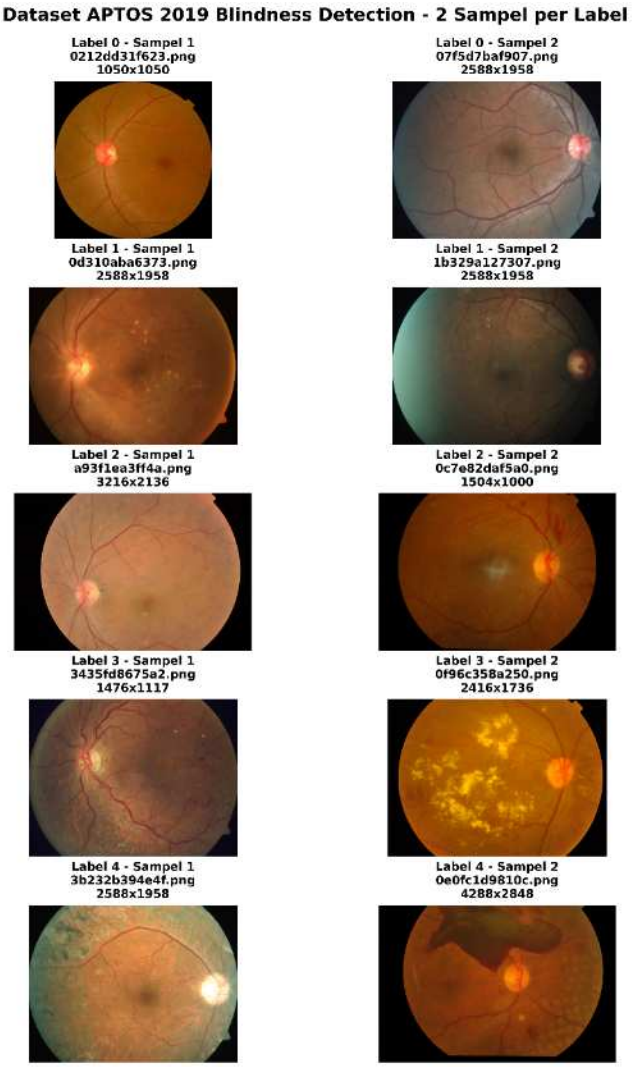
Penelitian ini memanfaatkan lingkungan komputasi menggunakan *Kaggle Notebook* yang didukung oleh akselerator GPU P100 untuk menjalankan seluruh kode pelatihan, modeling, hingga evaluasi model. Penelitian ini juga menggunakan *library* torch (PyTorch) sebagai *framework* utama untuk membangun, melatih, dan mengevaluasi model, serta timm (PyTorch Image Models) yang digunakan sebagai pembentuk model dan *backbone* untuk mengakses berbagai arsitektur *Vision Transformer* dan *EfficientNet* yang telah dilatih sebelumnya, maupun *library* lain seperti fvcore, scikit-learn, numpy, random, pandas, matplotlib, dan cv2 (OpenCV).

B. Deskripsi Dataset

Dataset yang digunakan adalah *Dataset APTOS 2019 Blindness Detection*, yang tersedia secara publik melalui platform Kaggle, terdiri atas 3662 gambar citra *fundus* retina yang digunakan untuk mendeteksi dan mengklasifikasikan tingkat keparahan Retinopati Diabetik. Dataset ini dikumpulkan dari *Aravind Eye Hospital* di India dan telah diperiksa serta dikategorikan oleh dokter mata berpengalaman [10]. Dataset ini terbagi ke dalam lima kelas tingkat keparahan Retinopati Diabetik berdasarkan *International Clinical Diabetic Retinopathy* (ICDR) [11].

- **0 (No DR)** - Tidak ada tanda-tanda retinopati diabetik. Berjumlah 1805 gambar.
- **1 (Mild DR)** - Terdapat sedikit mikroaneurisma (tonjolan kecil di pembuluh darah akibat gula darah tinggi) pada retina [12]. Berjumlah 370 gambar.
- **2 (Moderate DR)** - Ada peningkatan mikroaneurisma, perdarahan, serta kelainan pembuluh darah lain yang lebih menonjol, seperti perdarahan dan kelainan struktur vaskular [12]. Berjumlah 999 gambar.
- **3 (Severe DR)** - Banyak perdarahan, adanya IRMA (*intraretinal microvascular abnormalities* atau kelainan mikrovaskular intraretina) dan *venous beading* [12]. Berjumlah 193 gambar.
- **4 (Proliferative DR)** - Stadium paling lanjut, ditandai dengan neovaskularisasi (pertumbuhan pembuluh darah abnormal pada retina dan diskus optikus) [12]. Berjumlah 295 gambar.

Selain itu, resolusi citra dalam dataset ini bervariasi, mulai dari 474×358 piksel hingga 4288x2848 piksel. *Dataset APTOS 2019 Blindness Detection* juga dikenal memiliki distribusi kelas yang tidak seimbang, di mana kelas "No DR" memiliki jumlah gambar yang jauh lebih banyak dibandingkan kelas lainnya. Ketidakseimbangan ini berpotensi menimbulkan bias terhadap kelas mayoritas, sehingga pada penelitian ini penulis menerapkan teknik *offline data augmentation* untuk menambah jumlah sampel pada kelas minoritas agar lebih seimbang.



Gbr. 2 Sampel Citra *Fundus* Retinopati Diabetik per Label.

C. Preprocessing Data

Preprocessing Data merupakan langkah fundamental untuk mempersiapkan data citra mentah dengan meningkatkan kualitas dan keseragaman data input agar sesuai dengan input

model *deep learning* yang diinginkan serta mendukung stabilitas pelatihan model tersebut. Tujuan utama dari langkah pra-pemrosesan ini adalah untuk meningkatkan perilaku sistem, seperti menghilangkan derau, meningkatkan kontras, menormalisasi citra, yang pada akhirnya meningkatkan kualitas gambar demi mendapatkan kinerja yang lebih baik untuk algoritma deep learning tersebut [13].

Dalam penelitian ini, *preprocessing* yang dilakukan antara lain:

1. *Cropping*

Tahap cropping dilakukan untuk menghilangkan area hitam di sekitar citra fundus yang tidak memiliki informasi klinis penting, agar model fokus pada bagian utama retina serta meminimalisasi noise visual yang dapat mengganggu proses ekstraksi fitur.

2. *Resizing*

Resizing adalah tahap pra-pemrosesan yang krusial dalam *computer vision* [14] . Dalam penelitian ini, semua citra akan disesuaikan ukurannya mengikuti resolusi input yang direkomendasikan untuk setiap varian arsitektur, yaitu *EfficientNet* yang menggunakan prinsip *compound scaling*, maupun *Vision Transformer* yang juga memiliki dimensi input standar tersendiri. Selain itu, citra juga akan diubah ukurannya menjadi resolusi lebih tinggi, yaitu 512×512 piksel, untuk meminimalkan hilangnya informasi penting seperti detail halus pada citra fundus yang justru relevan dalam mendeteksi tingkat keparahan Retinopati Diabetik.

TABEL I
RESOLUSI INPUT YANG DISARANKAN PER MODEL

Model	Resolusi Input yang Disarankan
EfficientNet-B0	224 × 224
EfficientNet-B1	240 × 240
EfficientNet-B2	288 × 288
EfficientNet-B3	320 × 320
EfficientNet-B4	384 × 384
MobileViT-XS	256 × 256

3. CLAHE (*Contrast Limited Adaptive Histogram Equalization*)

Teknik ini digunakan untuk meningkatkan kontras lokal pada citra gambar tanpa menambah gangguan *noise* di area yang homogen, karena sangat bermanfaat terutama untuk citra *fundus* retina yang kerap mengalami variasi pencahayaan. Penelitian yang dilakukan oleh Owler et al. [15], CLAHE terbukti menjadi metode *preprocessing* paling efektif dibandingkan dengan metode lain seperti teknik N4 atau tanpa pemrosesan sama sekali [16].

4. Normalisasi

Normalisasi yang dilakukan pada penelitian ini adalah mengubah rentang nilai piksel ke skala standar *ImageNet* sehingga membantu mempercepat konvergensi model dan meningkatkan kinerja, dan terbukti sangat efektif untuk model *Vision Transformer* (ViT) maupun CNN karena memberikan manfaat penting dalam meningkatkan performanya masing-masing [17].

Penelitian ini diterapkan juga *offline data augmentation* setelah tahap *preprocessing* untuk menjaga keseragaman distribusi data asli sekaligus mengatasi ketidakseimbangan jumlah sampel antar kelas melalui transformasi seperti *rotation*, *flip*, penyesuaian kecerahan, *zoom*, dan penambahan *noise*. Jumlah sampel pada masing-masing kelas hasil augmentasi diseimbangkan dengan menyesuaikan ke kelas mayoritas, yaitu kelas “No DR” (label 0) yang memiliki 1.805 gambar. Dengan demikian, seluruh kelas (0–4) dipastikan memiliki 1.805 sampel, sehingga total dataset menjadi 9.025 citra dari yang sebelumnya 3662 citra.

D. Pembagian Dataset

Setelah *preprocessing*, *Dataset APTOS 2019 Blindness Detection* dibagi menjadi tiga subset, yaitu *train*, *validation*, dan *test*. Pendekatan penggunaan subset *validation* terpisah penting untuk dilakukan untuk memastikan evaluasi kinerja model yang konsisten, objektif, dan tidak bias, sehingga dapat mengurangi risiko *overfitting* model maupun membuat model lebih dapat diandalkan saat diaplikasikan di dunia nyata [18].

Rasio pembagian yang digunakan adalah 80% untuk *train set*, 10% untuk *validation set*, dan 10% untuk *test set*. Karena sudah dilakukan *offline augmentation* sebelumnya, maka berdasarkan total 9025 gambar dalam *dataset APTOS 2019*, jumlah gambar di setiap subset adalah sebagai berikut:

- Total gambar di *Train Set*: 7220 gambar
- Total gambar di *Validation Set*: 902 gambar
- Total gambar di *Test Set*: 903 gambar

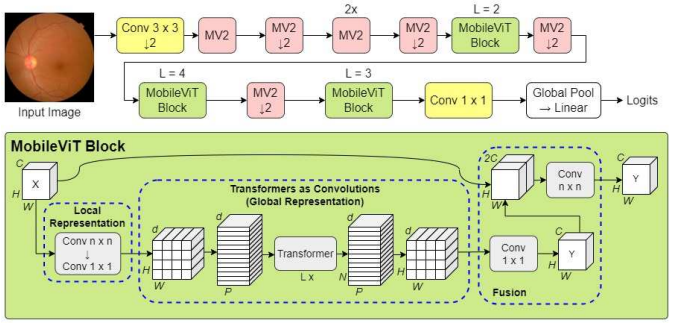
E. Hybrid Vision Transformer dan EfficientNet

Penelitian ini akan menguji berbagai kombinasi model *Vision Transformer* (ViT) ringan dan *EfficientNet* melalui strategi *hybrid* fitur. Pendekatan *hybrid* ini bertujuan untuk menggabungkan kemampuan ekstraksi fitur lokal dari CNN (*EfficientNet*) dengan kemampuan pemodelan konteks global dari ViT sehingga dapat menghasilkan representasi citra yang lebih komprehensif untuk klasifikasi level Retinopati Diabetik.

Arsitektur-arsitektur tersebut antara lain:

1. MobileViT-XS

MobileViT adalah ViT ringan yang menggabungkan kekuatan CNN dan ViT untuk tugas spesifik *mobile vision*. MobileViT yang digunakan dalam penelitian ini juga telah dilatih sebelumnya (*pretrained*) dengan *dataset ImageNet-1k*. MobileViT-XS memiliki sekitar 2.3 juta parameter dengan ~1.1 GMACs pada input resolusi 256 x 256 piksel, yang menjadikannya sangat efisien untuk perangkat *mobile* di mana sumber daya dan *bandwidth* jaringan terbatas.



Gbr. 3 Arsitektur MobileViT.

2. *EfficientNet*

EfficientNet adalah keluarga arsitektur CNN yang dikenal karena efisiensi dan akurasinya yang sangat baik dengan mengombinasikan *compound scaling* yaitu menskalakan kedalaman, lebar, dan resolusi secara terpadu dengan satu koefisien [19]. Perbedaan antara varian *EfficientNet*-B0 hingga *EfficientNet*-B4 sesuai dengan dokumentasi PyTorch adalah sebagai berikut.

a. *EfficientNet*-B0

Merupakan *baseline* yang menjadi titik awal perhitungan *compound scaling*, dengan menggunakan resolusi input 224×224 piksel, sekitar 5,3 juta parameter, dan ~0,39 GFLOPs.

b. *EfficientNet*-B1

Memiliki sedikit peningkatan kedalaman jaringan dan lebar kanal, dengan menggunakan resolusi input 240×240 piksel, sekitar 7,8 juta parameter, dan ~0,69 GFLOPs.

c. *EfficientNet*-B2

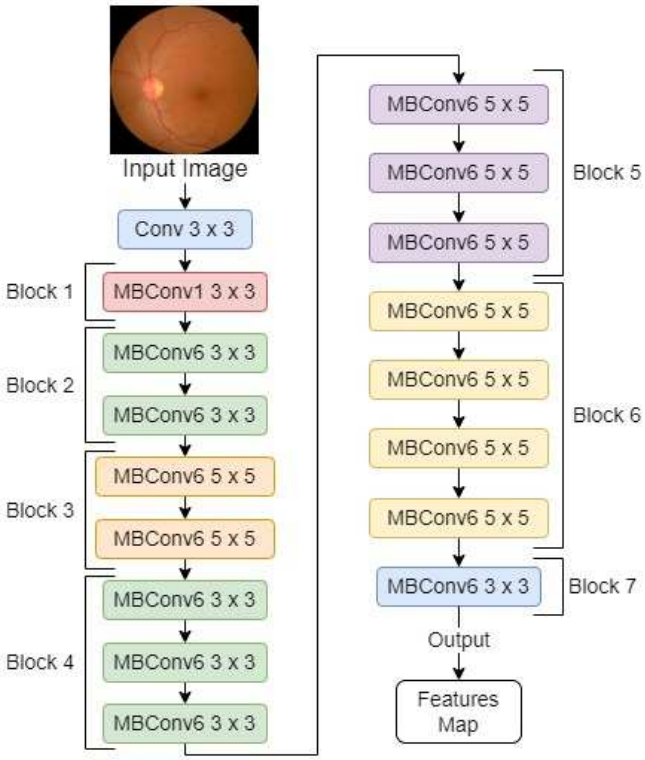
Memiliki peningkatan kedalaman, lebar, dan resolusi input ke 288×288 piksel, dengan jumlah parameter ~9,1 juta dan beban komputasi ~1,09 GFLOPs.

d. *EfficientNet*-B3

Menggunakan resolusi input yang lebih besar, 320×320 piksel, dengan jumlah parameter sekitar 12,2 juta dan ~1,83 GFLOPs.

e. *EfficientNet*-B4

Salah satu varian yang cukup besar dengan resolusi input 384×384 piksel, parameter ~19,3 juta, dan beban komputasi ~4,39 GFLOPs.



Gbr. 4 Arsitektur *EfficientNet*-B0.

F. Pengukuran Kinerja Model

Untuk mengevaluasi efektivitas dan efisiensi model *hybrid* yang diusulkan, beberapa metrik kinerja akan diukur secara komprehensif. Metrik-metrik ini akan memberikan gambaran lengkap tentang kemampuan model dalam mengklasifikasikan level Retinopati Diabetik dan seberapa efisien model tersebut beroperasi.

1. Akurasi (*Accuracy*)

Proporsi prediksi yang benar dari total prediksi yang menggambarkan perbandingan jumlah prediksi yang tepat dengan total keseluruhan prediksi, dengan rumus Persamaan 1 sebagai berikut [20].

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

2. Presisi (*Precision*)

Rasio *true positive* terhadap total hasil positif yang diprediksi (*true positive* + *false positive*), mengukur seberapa relevan hasil positif yang diprediksi oleh model. Rumus presisi adalah sebagai berikut dalam Persamaan 2 [20].

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

3. *Recall* (Sensitivitas)

Rasio *true positive* terhadap total positif aktual (*true positive* + *false negative*) yang mengukur kemampuan model untuk menemukan semua instansi positif. Rumus *recall* adalah sebagai berikut dalam Persamaan 3 [20].

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

4. *F1-Score*

Rata-rata harmonik dari presisi dan *recall* yang sangat berguna ketika ada ketidakseimbangan antara kedua metrik tersebut. Contohnya, pada sebuah tugas di mana kesalahan *false positives* (FP) dan *false negatives* (FN) memiliki dampak yang sama pentingnya, *f1-score* mampu menyajikan gambaran kinerja model yang lebih seimbang dibandingkan hanya melihat presisi atau *recall* secara terpisah. Persamaan *f1-score* ditunjukkan dalam Persamaan 4 sebagai berikut [20].

$$F1\ score = 2 * \frac{TP}{TP + FP + FN} \quad (4)$$

Selain metrik efektivitas seperti akurasi, presisi, *recall*, dan *f1-score*, penelitian ini juga memperhatikan metrik efisiensi model yang mencakup aspek-aspek berikut:

1. Waktu Pelatihan (*Training Time*)

Durasi yang dibutuhkan untuk melatih model *deep learning* dari awal hingga konvergensi atau mencapai kinerja yang diinginkan, yang diukur dalam satuan waktu (menit dan detik) dan mencerminkan biaya komputasi untuk mengembangkan model.

2. FLOPs (*Floating-Point Operations*)

Ukuran kompleksitas komputasi sebuah model, yang menghitung total operasi *floating-point* yang dibutuhkan selama proses *training*, yang diukur dalam satuan GFLOPs (miliar operasi) dan menunjukkan seberapa berat beban komputasi model tersebut.

3. Ukuran Model (*Model Size*)

Merupakan ukuran model yang telah dilatih dalam *Megabyte* (MB) saat disimpan di *disk* atau dimuat ke memori. Hal ini

sangat penting terutama saat model diterapkan pada perangkat dengan memori terbatas, seperti perangkat *mobile*.

III. HASIL DAN PEMBAHASAN

A. Hasil Resize Vision Transformer

TABEL II
HASIL EFEKTIVITAS *RESIZE VISION TRANSFORMER* DENGAN *BACKBONE* MOBILEVIT-XS

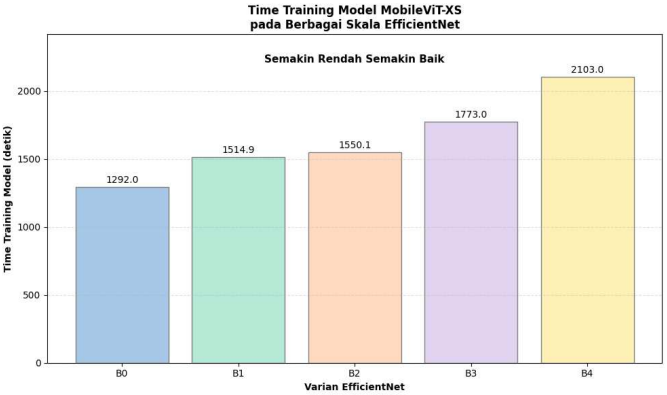
XS	***Train Acc***	***Val Acc***	***Val Prec***	***Val Rec***	***Val F1-S***
B0	94.16%	89.58%	90.10%	89.58%	89.66%
B1	**97.76%**	**92.79%**	**92.90%**	**92.79%**	**92.79%**
B2	95.22%	90.35%	90.61%	90.36%	90.41%
B3	96.16%	90.24%	90.30%	90.25%	90.27%
B4	92.76%	89.80%	90.06%	89.81%	89.81%
XS	***Test Acc***	***Test Prec***	***Test Rec***	***Test F1-S***	
B0	89.15%	89.60%	89.15%	89.17%	
B1	91.47%	91.60%	91.48%	91.38%	
B2	90.59%	90.63%	90.59%	90.60%	
B3	91.14%	91.24%	91.14%	91.16%	
B4	**91.58%**	**91.80%**	**91.58%**	**91.58%**	

Pada Tabel II, hasil efektivitas *backbone* MobileViT-XS terbaik adalah antara *hyrbid EfficientNet*-B1 hingga *EfficientNet*-B4. Sedangkan hasil terburuk didapatkan oleh *EfficientNet*-B0 dengan nilai metrik evaluasi *validation* dan *test*-nya tidak menyentuh angka 90%. *Train accuracy* dan *validation accuracy* tertinggi adalah pada varian *EfficientNet*-B1, masing-masing sebesar 97,76% dan 92,79%, sedangkan *test accuracy* terbaik diraih oleh varian *EfficientNet*-B4 sebesar 91,58%.

Berdasarkan hasil pengujian eksperimen pertama ini, terlihat bahwa kombinasi dengan *EfficientNet*-B1 hingga *EfficientNet*-B4 cenderung memberikan kinerja efektivitas terbaik, sementara pada varian *EfficientNet*-B0 menunjukkan hasil yang cenderung relatif lebih rendah. Hal ini terjadi karena pada proses *resize* citra menjadi ukuran sesuai *Vision Transformer* (ViT), yaitu 256×256 piksel untuk MobileViT-XS menunjukkan bahwa hasil pelatihan yang paling optimal ketika menggunakan varian *EfficientNet*-B1. Varian ini memiliki ukuran input dasar 240×240 piksel yang paling mendekati resolusi input ViT, sehingga mampu menghasilkan keseimbangan yang lebih baik antara ekstraksi fitur lokal dan representasi global. Selain itu, *EfficientNet*-B2 hingga *EfficientNet*-B4 juga memperlihatkan pola performa yang sejalan dengan *EfficientNet*-B1, meskipun dengan tingkat efektivitas yang sedikit lebih rendah. Kesesuaian skala ini berkaitan dengan konsep *compound scaling* pada *EfficientNet*, yaitu peningkatan skala jaringan dilakukan secara seimbang pada kedalaman (*depth*), lebar (*width*), dan resolusi input, sehingga varian *EfficientNet*-B1 hingga *EfficientNet*-B4 tetap mampu menjaga stabilitas dan efektivitas pelatihan pada konfigurasi resolusi yang digunakan.

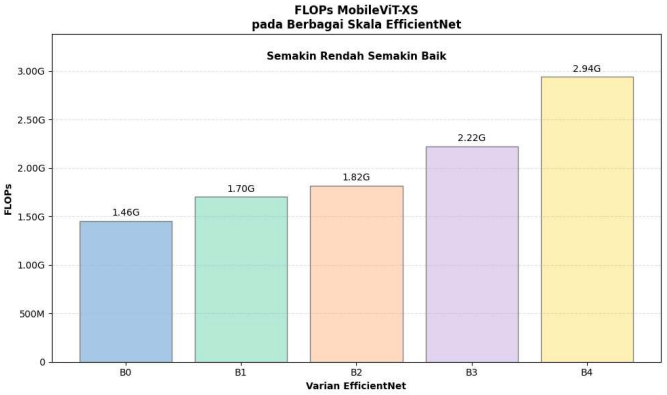
Sebaliknya, penggunaan *EfficientNet*-B0 dengan resolusi dasar 224x224 piksel menghasilkan performa yang kurang optimal. Hal ini disebabkan oleh perbedaan skala yang menyebabkan *Vision Transformer* lebih dominan dalam mengekstraksi fitur global, sementara informasi lokal yang penting menjadi berkurang. Dengan demikian, kombinasi ukuran input yang paling sepadan antara *Vision Transformer* dan *EfficientNet* terletak pada varian *EfficientNet*-B1, di mana keseimbangan antara fitur lokal dan global dapat tercapai dengan lebih baik.

Berikutnya adalah hasil efisiensi dari setiap varian *Vision Transformer* yang dikombinasikan dengan *EfficientNet* dengan menggunakan *resize* sesuai spesifikasi *Vision Transformer* yaitu MobileViT-XS dengan 256x256 piksel, *batch size* 16, *learning rate* 1e-4, *optimizer* AdamW, dan pada 15 *epoch* disajikan dalam Gambar 5, Gambar 6, dan Gambar 7.



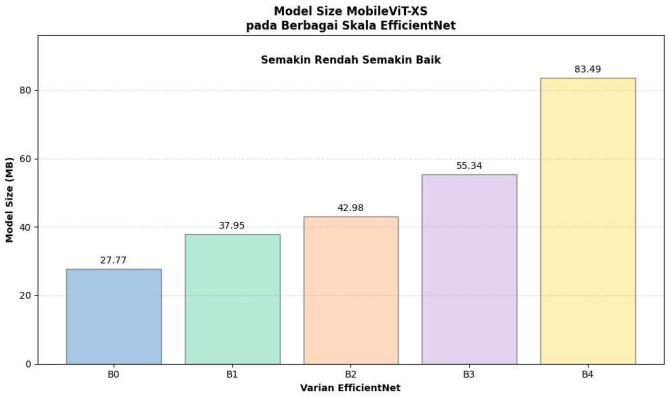
Gbr. 5 *Time Training Model* sesuai *resize* ViT.

Berdasarkan Gambar 5, lamanya *Time Training Model* selaras dengan meningkatnya varian *EfficientNet* yang digunakan. *Backbone* MobileViT-XS mencatat waktu pelatihan yang meningkat dari 21 menit 32,02 detik pada *EfficientNet*-B0 menjadi 35 menit 3,01 detik pada *EfficientNet*-B4.



Gbr. 6 FLOPs sesuai *resize* ViT.

Berdasarkan Gambar 6, peningkatan varian *EfficientNet* yang digunakan berbanding lurus dengan peningkatan nilai FLOPs yang dihasilkan. Pada *backbone* MobileViT-XS mencatat rentang FLOPs dari sekitar 1,46 GFLOPs pada *EfficientNet*-B0 hingga 2,94 GFLOPs pada *EfficientNet*-B4.



Gbr. 7 *Model Size* sesuai *resize* ViT.

Berdasarkan Gambar 7, peningkatan varian *EfficientNet* yang digunakan berbanding lurus dengan kenaikan *Model Size*. Pada MobileViT-XS ukuran model meningkat dari 27,77 MB pada *EfficientNet*-B0 hingga 83,49 MB pada *EfficientNet*-B4.

B. Hasil Resize Compound Scaling

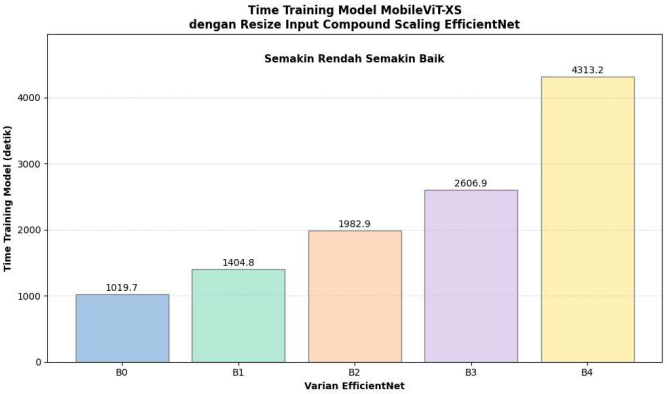
TABEL III
HASIL EFEKTIVITAS *RESIZE COMPOUND SCALING* DENGAN *BACKBONE* MOBILEVIT-XS

XS	***Train Acc***	***Val Acc***	***Val Prec***	***Val Rec***	***Val F1-S***
B0	93.86%	89.02%	89.36%	89.03%	89.08%
B1	**97.38%**	**92.57%**	**92.66%**	**92.58%**	**92.58%**
B2	95.82%	91.46%	91.52%	91.47%	91.46%
B3	97.23%	91.80%	91.85%	91.80%	91.80%
B4	94.03%	88.58%	88.58%	88.59%	88.54%
XS	***Test Acc***	***Test Prec***	***Test Rec***	***Test F1-S***	
B0	88.04%	88.23%	88.04%	87.98%	
B1	**91.92%**	**91.97%**	**91.91%**	**91.91%**	
B2	90.03%	90.13%	90.03%	90.06%	
B3	90.48%	90.45%	90.47%	90.46%	
B4	90.37%	90.67%	90.36%	90.38%	

Pada Tabel III, hasil efektivitas *backbone* MobileViT-XS terbaik adalah antara *hybrid EfficientNet*-B1 hingga *EfficientNet*-B3. Sementara itu, hasil kinerja terburuk terlihat pada *EfficientNet*-B0 dan *EfficientNet*-B4 dengan nilai metrik evaluasi pada data validasi ataupun data uji yang tidak menyentuh angka 90%. *Train accuracy, validation accuracy,* dan *test accuracy* terbaik adalah pada *EfficientNet*-B1, masing-masing sebesar 97,38%, 92,57%, dan 91,92%.

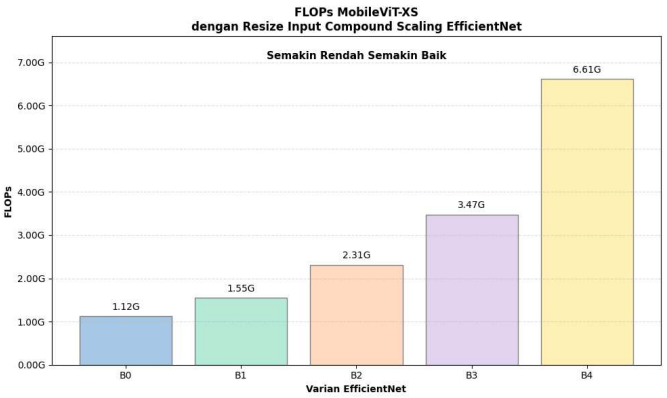
Berdasarkan hasil pengujian eksperimen kedua, terlihat bahwa pola performa yang dihasilkan tidak jauh berbeda dengan eksperimen pertama, bahwa secara konsisten *hybrid* dengan *EfficientNet*-B1 hingga *EfficientNet*-B3 menunjukkan hasil efektivitas terbaik, sedangkan *EfficientNet*-B0 dan *EfficientNet*-B4 menghasilkan kinerja yang relatif lebih rendah. Kemiripan antara hasil eksperimen pertama dan eksperimen kedua mengindikasikan bahwa keseimbangan skala resolusi input terhadap kemampuan arsitektur model atau *backbone* dalam memproses dan mengintegrasikan informasi visual memiliki peran yang lebih dominan dibandingkan sekadar strategi *resize* itu sendiri dalam menyeragamkan dimensi input. Meskipun pada eksperimen ini resolusi input telah disesuaikan secara eksplisit dengan prinsip *compound scaling*, hasil yang diperoleh tetap menunjukkan bahwa varian *EfficientNet*-B1 hingga *EfficientNet*-B3 tetap lebih mampu menjaga keseimbangan antara kompleksitas fitur lokal yang diekstraksi olehnya dan kemampuan *Vision Transformer* dalam menangkap representasi global. Sehingga, hal ini menjelaskan mengapa peningkatan resolusi input tidak selalu berbanding lurus dengan peningkatan performa, terutama ketika kapasitas representasi dan arsitektur model tidak lagi berada pada titik keseimbangan yang optimal. Dengan demikian, hasil eksperimen ini menegaskan bahwa varian *EfficientNet*-B1 hingga *EfficientNet*-B3 merupakan titik *trade-off* optimal dalam metode *hybrid Vision Transformer* dan *EfficientNet*, bahkan ketika resolusi input disesuaikan sepenuhnya dengan prinsip *compound scaling*.

Selanjutnya adalah hasil efisiensi dari setiap varian *Vision Transformer* yang dikombinasikan dengan *EfficientNet* dengan menggunakan *resize* berbasis *compound scaling*, dengan *batch size* 16, *learning rate* 1e-4, *optimizer* AdamW, dan pada 15 *epoch* yang disajikan dalam Gambar 8, Gambar 9, dan Gambar 10.



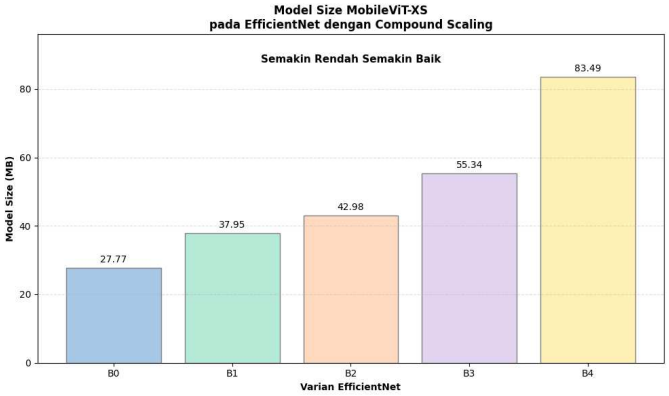
Gbr. 8 *Time Training Model* sesuai *resize Compound Scaling*.

Berdasarkan Gambar 8, lamanya *Time Training Model* selaras dengan meningkatnya varian *EfficientNet* yang digunakan. Peningkatan waktu pelatihan dari varian *EfficientNet*-B0 hingga *EfficientNet*-B4 menunjukkan tren yang lebih tajam dibandingkan dengan skenario *resize* sesuai spesifikasi *Vision Transformer* pada eksperimen pertama. *Backbone* MobileViT-XS mencatat waktu pelatihan yang meningkat dari 16 menit 59,69 detik pada *EfficientNet*-B0 menjadi 71 menit 53,21 detik pada *EfficientNet*-B4.



Gbr. 9 FLOPs sesuai *resize Compound Scaling*.

Berdasarkan Gambar 9, yang menunjukkan hasil *Floating-Point Operations Per Second* (FLOPs), tercatat bahwa MobileViT-XS memiliki FLOPs yang meningkat dari 1,12 GFLOPs pada *EfficientNet*-B0 menjadi 6,61 GFLOPs pada *EfficientNet*-B4. Selain itu, peningkatan varian *EfficientNet* dari *EfficientNet*-B0 hingga *EfficientNet*-B4 berbanding lurus dengan meningkatnya jumlah FLOPs yang dihasilkan, dengan tren peningkatan yang lebih tajam dibandingkan skenario *resize* sesuai spesifikasi *Vision Transformer* pada eksperimen pertama.



Gbr. 10 *Model Size* sesuai *resize Compound Scaling*.

Sebagai metrik efisiensi terakhir, *Model Size* ditunjukkan pada Gambar 10. Hasil pengujian menunjukkan bahwa *backbone* MobileViT-XS mencatat ukuran model yang meningkat dari 27,77 MB pada *EfficientNet*-B0 hingga 83,49 MB pada *EfficientNet*-B4. Selain itu, penggunaan varian *EfficientNet* yang lebih tinggi berbanding lurus dengan meningkatnya *Model Size* yang dihasilkan. Menariknya, nilai *Model Size* pada eksperimen kedua tidak mengalami perubahan dibandingkan dengan eksperimen pertama yang menerapkan *resize* sesuai spesifikasi *Vision Transformer*, sehingga menghasilkan ukuran model yang identik.

C. Hasil Resize 512x512 piksel

TABEL IV
HASIL EFEKTIVITAS *RESIZE 512X512 PIKSEL* DENGAN *BACKBONE* MOBILEVIT-XS

XS	***Train Acc***	***Val Acc***	***Val Prec***	***Val Rec***	***Val F1-S***
B0	95.65%	90.35%	90.35%	90.36%	90.33%
B1	97.92%	91.80%	91.87%	91.80%	91.80%
B2	97.55%	**93.02%**	**93.04%**	**93.02%**	**93.00%**
B3	**98.38%**	91.80%	91.85%	91.80%	91.76%
B4	94.70%	90.35%	90.80%	90.35%	90.44%
XS	***Test Acc***	***Test Prec***	***Test Rec***	***Test F1-S***	
B0	92.69%	92.76%	92.69%	92.67%	
B1	**93.24%**	**93.28%**	**93.24%**	**93.25%**	
B2	92.47%	92.51%	92.47%	92.48%	
B3	92.47%	92.59%	92.47%	92.51%	
B4	89.92%	90.18%	89.92%	89.95%	

Pada Tabel IV, hasil efektivitas terbaik untuk *backbone* MobileViT-XS adalah antara *hybrid* dengan *EfficientNet*-B0 hingga *EfficientNet*-B3. Dan hasil terburuk pada *EfficientNet*-B4 yang memiliki nilai metrik evaluasi pada data uji yang tidak menyentuh angka 90%. *Train accuracy* tertinggi adalah pada varian *EfficientNet*-B3 sebesar 98,38%, *validation accuracy* tertinggi pada *EfficientNet*-B2 sebesar 93.02%, dan *test accuracy* terbaik adalah pada *EfficientNet*-B1 sebesar 93,24%.

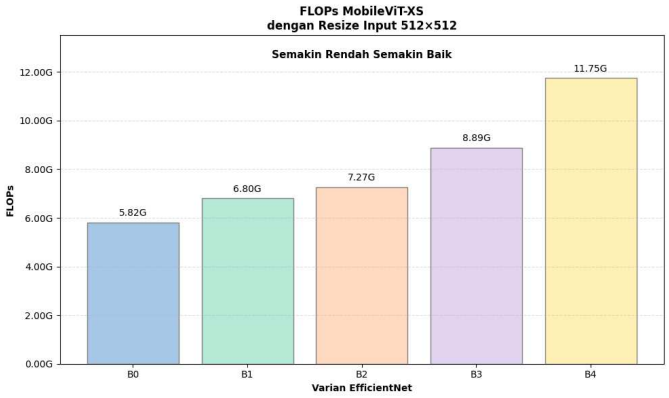
Pola yang terlihat pada *backbone* MobileViT-XS adalah dengan performa terbaik yang terkonsentrasi pada *EfficientNet*-B1 hingga *EfficientNet*-B3. Meskipun *EfficientNet*-B3 unggul pada *train accuracy* karena memiliki kapasitas yang lebih besar untuk mempelajari pola pada pelatihan data, *EfficientNet*-B1 tetap menunjukkan performa terbaik pada *test accuracy* karena kemampuan generalisasi yang lebih baik terhadap data yang belum pernah dilihat.

Secara keseluruhan, eksperimen ini menegaskan bahwa penggunaan resolusi input tinggi 512×512 piksel tidak secara otomatis meningkatkan kinerja seluruh kombinasi model. Efektivitasnya sangat bergantung pada keseimbangan antara resolusi input, kapasitas *EfficientNet*, dan kemampuan *backbone Vision Transformer* dalam mengelola kompleksitas fitur. Varian *EfficientNet*-B1 hingga *EfficientNet*-B3 kembali terbukti sebagai konfigurasi yang paling stabil, sementara *EfficientNet*-B4 cenderung mengalami degradasi kinerja akibat kompleksitas fitur yang tidak sebanding dengan kemampuan representasi global model *Vision Transformer* yang digunakan.

Selanjutnya, adalah hasil efisiensi dari setiap varian *Vision Transformer* yang dikombinasikan dengan *EfficientNet* dengan menggunakan resize 512x512 piksel, *batch size* 8, *learning rate* 1e-4, *optimizer* AdamW, dan pada 15 *epoch* yang disajikan dalam Gambar 11, Gambar 12, dan Gambar 13.

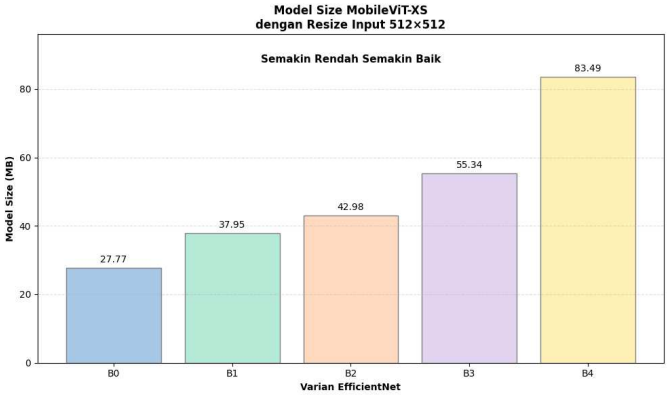
Gbr. 11 *Time Training Model* sesuai *resize* 512x512 piksel.

Pada Gambar 11 terlihat bahwa hasil *Time Training Model* menunjukkan pola peningkatan varian *EfficientNet* yang digunakan berbanding lurus dengan meningkatnya waktu pelatihan model. *Backbone* MobileViT-XS mencatat waktu dari 89 menit 0,61 detik pada *EfficientNet*-B0 hingga 138 menit 17,05 detik pada *EfficientNet*-B4.



Gbr. 12 FLOPs sesuai *resize* 512x512 piksel.

Berdasarkan Gambar 12, metrik *Floating-Point Operations Per Second* (FLOPs) menunjukkan bahwa MobileViT-XS mencatat FLOPs yang meningkat dari 5,82 GFLOPs pada *EfficientNet*-B0 menjadi 11,75 GFLOPs pada *EfficientNet*-B4. Temuan ini menegaskan bahwa penggunaan varian *EfficientNet* serta penambahan resolusi input menjadi 512×512 piksel dapat memberikan kontribusi yang semakin signifikan terhadap total FLOPs.



Gbr. 13 *Model Size* sesuai *resize* 512x512 piksel.

Sebagai metrik efisiensi terakhir, yaitu *Model Size*, pada Gambar 13 terlihat bahwa MobileViT-XS mencatat ukuran model yang meningkat dari 27,77 MB pada *EfficientNet*-B0 hingga 83,49 MB pada *EfficientNet*-B4. Penggunaan varian *EfficientNet* yang semakin tinggi juga berbanding lurus dengan meningkatnya *Model Size* yang dihasilkan. Menariknya, pada eksperimen ketiga ini, nilai *Model Size* tidak mengalami perubahan dibandingkan dengan eksperimen pertama dan kedua yang menerapkan *resize* sesuai spesifikasi *Vision Transformer* dan prinsip *compound scaling*, sehingga menghasilkan ukuran model yang identik.

IV. KESIMPULAN

Berdasarkan serangkaian eksperimen komprehensif melalui *resize* sesuai spesifikasi *Vision Transformer*, *resize* sesuai prinsip *compound scaling*, dan *resize* 512x512 piksel, dapat disimpulkan bahwa *hybrid* antara MobileViT-XS dan *EfficientNet*-B1 dengan *resize* 512×512 piksel, *batch size* 8, *learning rate* 1e-4, *optimizer* AdamW, dan 15 *epoch* memberikan kinerja paling optimal. Model *hybrid* ini tidak hanya menghasilkan efektivitas melalui *validation accuracy* dan *test accuracy* tertinggi dibandingkan dengan konfigurasi lain yaitu masing-masing sebesar 91,80% dan 93,24%, tetapi juga menunjukkan stabilitas kinerja yang baik dengan efisiensi komputasi yang masih masuk akal. Temuan ini menegaskan bahwa *hybrid Vision Transformer* dan *EfficientNet* mampu memanfaatkan keunggulan masing-masing arsitektur secara komplementer, sehingga menghasilkan model klasifikasi retinopati diabetik yang efektif dan efisien dalam sistem deteksi dini berbasis kecerdasan buatan.

UCAPAN TERIMA KASIH

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa, karena atas rahmat dan karunia-Nya penulis dapat menyelesaikan penelitian ini dengan baik. Penulis juga menyampaikan apresiasi dan rasa terima kasih yang sebesar-besarnya kepada seluruh teman-teman yang telah berkontribusi dan terlibat dalam pelaksanaan penelitian ini, kepada dosen pembimbing yang dengan penuh dedikasi memberikan arahan, masukan, dan pendampingan sejak tahap awal hingga penyelesaian penelitian, serta kepada orang tua yang senantiasa memberikan dukungan moral dan doa tanpa pamrih sepanjang proses studi dan penelitian berlangsung.

REFERENSI

[1] D. Ghosh Aronno and S. Saeha, “Diabetic Retinopathy Detection Using CNN with Residual Block and DCGAN,” *arXiv (Cornell University)*, vol. 1, Jan. 2025, doi: https://doi.org/10.48550/arXiv.2501.02300.

[2] H. Sun *et al.*, “IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045,” *Diabetes Res. Clin. Pract.*, vol. 183, p. 109119, 2022, doi: 10.1016/j.diabres.2021.109119.

[3] T. Y. Wong and T. E. Tan, “The Diabetic Retinopathy ‘Pandemic’ and Evolving Global Strategies: The 2023 Friedenwald Lecture,” *Invest. Ophthalmol. Vis. Sci.*, vol. 64, no. 15, 2023, doi: 10.1167/iovs.64.15.47.

[4] W. Zhang, V. Belcheva, and T. Ermakova, “Interpretable Deep Learning for Diabetic Retinopathy: A Comparative Study of CNN, ViT, and Hybrid Architectures,” *Computers*, vol. 14, no. 5, p. 187, May 2025, doi: 10.3390/computers14050187.

[5] T. Karkera, C. Adak, S. Chattopadhyay, and M. Saqib, “Detecting severity of Diabetic Retinopathy from fundus images: A transformer network-based review,” *Neurocomputing*, vol. 597, pp. 1–28, 2024, doi: 10.1016/j.neucom.2024.127991.

[6] S. Takahashi *et al.*, “Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review,” *J. Med. Syst.*, vol. 48, no. 1, pp. 1–22, 2024, doi: 10.1007/s10916-024-02105-8.

[7] J. Rautaray *et al.*, “Leveraging FastViT based knowledge distillation with EfficientNet-B0 for diabetic retinopathy severity classification,” *SLAS Technol.*, vol. 33, no. June, p. 100325, 2025, doi: 10.1016/j.slast.2025.100325.

[8] Y. Fu, Y. Ju, and D. Zhang, “MSEF-Net: A multi-scale EfficientNet Fusion for Diabetic Retinopathy grading,” *Biomed. Signal Process. Control*, vol. 98, no. November 2023, p. 106714, 2024, doi: 10.1016/j.bspc.2024.106714.

[9] V. Tanwar, B. Sharma, D. P. Yadav, and A. Mehbodniya, “Hybrid deep learning framework based on EfficientViT for classification of gastrointestinal diseases,” *Sci. Rep.*, vol. 15, no. 1, pp. 1–24, 2025, doi: 10.1038/s41598-025-12128-x.

[10] Karthik, Maggie, and S. Dane, “APTOS 2019 Blindness Detection,” 2019.

[11] H. Esmaeilkhanian *et al.*, "The relationship of diabetic retinopathy severity scales with frequency and surface area of diabetic retinopathy lesions," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 261, no. 11, pp. 3165–3176, 2023, doi: 10.1007/s00417-023-06145-7.

[12] S. Srinivasan *et al.*, "Inter-observer agreement in grading severity of diabetic retinopathy in wide-field fundus photographs," *Eye (Basingstoke)*, vol. 37, no. 6, pp. 1231–1235, 2023, doi: 10.1038/s41433-022-02107-1.

[13] D. Murcia-Gómez, I. Rojas-Valenzuela, and O. Valenzuela, "Impact of Image Preprocessing Methods and Deep Learning Models for Classifying Histopathological Breast Cancer Images," *Applied Sciences (Switzerland)*, vol. 12, no. 22, 2022, doi: 10.3390/app122211375.

[14] S. Saponara and A. Elhanashi, "Impact of Image Resizing on Deep Learning Detectors for Training Time and Model Performance," *Lecture Notes in Electrical Engineering*, vol. 866 LNEE, no. April, pp. 10–17, 2022, doi: 10.1007/978-3-030-95498-7_2.

[15] M. Youldash *et al.*, "Early Detection and Classification of Diabetic Retinopathy: A Deep Learning Approach," *AI (Switzerland)*, vol. 5, no. 4, pp. 2586–2617, 2024, doi: 10.3390/ai5040125.

[16] J. Owler and P. Rockett, "Influence of background preprocessing on the performance of deep learning retinal vessel detection," *Journal of Medical Imaging*, vol. 8, no. 06, 2021, doi: 10.1117/1.jmi.8.6.064001.

[17] G. L. Baroni, L. Rasotto, K. Roitero, A. Tulisso, C. Di Loreto, and V. Della Mea, "Optimizing Vision Transformers for Histopathology: Pretraining and Normalization in Breast Cancer Classification," *J. Imaging*, vol. 10, no. 5, 2024, doi: 10.3390/jimaging10050108.

[18] J. Levman, B. Ewenson, J. Apaloo, D. Berger, and P. N. Tyrrell, "Error Consistency for Machine Learning Evaluation and Validation with Application to Biomedical Diagnostics," *Diagnostics*, vol. 13, no. 7, 2023, doi: 10.3390/diagnostics13071315.

[19] M. Azmoodeh-Kalati, H. Shabani, M. S. Maghareh, Z. Barzegar, and R. Lashgari, "Leveraging an ensemble of EfficientNetV1 and EfficientNetV2 models for classification and interpretation of breast cancer histopathology images," *Sci. Rep.*, vol. 15, no. 1, pp. 1–25, Dec. 2025, doi: 10.1038/s41598-025-06853-6.

[20] S. A. Hicks *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Sci. Rep.*, vol. 12, no. 1, pp. 1–9, 2022, doi: 10.1038/s41598-022-09954-8.