Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 737~746
e-ISSN: 2548-6861

Comparative Analysis of MobileNetV3 and EfficientNetv2B0 in BISINDO Hand Sign Recognition Using MediaPipe Landmarks

Alief Khairul Fadzli [1]*, Majid Rahardi [2]*
* Informatics, Universitas Amikom Yogyakarta
aliefkhairulfadzli@students.amikom.ac.id [1], majid@amikom.ac.id [2]

Article Info

Article history:

Received 2025-11-29
Revised 2026-01-13
Accepted 2026-01-30

Keyword:

Sign Language Recognition, EfficientNetV2B0, MobileNetV3, MediaPipe, Deep Learning

ABSTRACT

Sign language is a vital communication medium for individuals with hearing and speech impairments. In Indonesia, more than 2.6 million people experience hearing disabilities, most of whom rely on Bahasa Isyarat Indonesia BISINDO for daily interaction. However, limited public understanding and the scarcity of professional interpreters continue to hinder inclusive communication. Recent advancements in *computer vision* and *deep learning* have enabled camera-based sign language recognition systems that are more affordable and practical compared to sensor-glove solutions. this study presents a comparative analysis between EfficientNetV2-B0 and MobileNetV3-Large in recognizing BISINDO hand sign alphabets using MediaPipe landmarks. The dataset was derived from BISINDO video recordings, from which hand landmarks were extracted using MediaPipe Hands and subsequently converted into two-dimensional skeletal images. In total, 10,309 skeletal images representing BISINDO alphabets A–Z were generated and used for model training and evaluation. Both models were trained under identical configurations using TensorFlow. The results show that MobileNetV3-Large achieved 89.67% test accuracy and an F1-score of 89.76%**,** while EfficientNetV2-B0 obtains 95.98% test accuracy and an F1-score of 95.93%. These findings highlight the trade-off between the higher classification accuracy of EfficientNetV2-B0 and the superior computational efficiency of MobileNetV3-Large. This research contributes to the development of lightweight, high-performance BISINDO recognition systems for assistive communication applications.

I. PENDAHULUAN

Komunikasi merupakan aspek fundamental dalam kehidupan manusia yang memungkinkan pertukaran gagasan, informasi, serta emosi. Bagi orang yang tuli atau kesulitan berbicara, Bahasa isyarat merupakan sarana dasar komunikasi. Melalui penggunaan terkoordinasi Gerakan tangan, posisi jari, dengarkan fisik, Bahasa isyarat menyampaikan makna sehingga memfasilitasi interaksi visual [1]. Menurut laporan survey global tahun 2023, 16% dari populasi dunia Adalah penyandang disabilitas dengan 80% totalnya tinggal di belahan bumi Selatan. Berdasarkan data Badan Pusat Statistik (BPS) tahun 2023, jumlah penyandang disabilitas Indonesia tercatat sebanyak 22,79 juta orang atau sekitar 8,5% dari total populasi Indonesia. Meskipun demikian, keterbatasan penerjemah dan rendahnya kesadaran masyarakat umum terhadap bahasa isyarat masih menjadi tantangan dalam menciptakan interaksi yang inklusif [2].

Perkembangan teknologi *computer vision* dan *deep learning* telah membuka peluang besar dalam otomatisasi pengenalan bahasa isyarat. *Convolutional Neural Network* (CNN) menjadi salah satu arsitektur paling populer karena kemampuannya mengekstraksi fitur spasial dari citra secara efektif dan efisien. Penelitian yang dilakukan oleh R. Sutjiadi *et al.* berhasil mengimplementasikan CNN untuk pengenalan huruf BISINDO, Dari hasil eksperimen dapat disimpulkan bahwa akurasi rata-rata mencapai 75,38% [3]. CNN juga memiliki keunggulan karena tidak membutuhkan perangkat keras tambahan seperti *data gloves*, sehingga lebih praktis digunakan [4].

Salah satu kemajuan signifikan dalam bidang pengenalan gesture tangan adalah penggunaan MediaPipe, kerangka kerja

yang dikembangkan oleh Google. MediaPipe mampu mendeteksi 21 titik landmark tangan dari citra RGB secara real-time maupun video dan menghasilkan koordinat dua dimensi yang stabil terhadap pencahayaan dan latar belakang [5]. Pendekatan ini mengubah data citra mentah menjadi representasi numerik skeletal yang memungkinkan model CNN fokus pada bentuk dan konfigurasi tangan tanpa terganggu oleh elemen non-esensial [6].

Namun demikian, efisiensi model tetap menjadi tantangan utama. Model ringan seperti MobileNetV3 dirancang untuk mengatasi masalah ini dengan menggabungkan *depthwise separable convolution*, blok *Squeeze-and-Excitation*, dan fungsi aktivasi *h-swish* yang meningkatkan efisiensi tanpa mengorbankan akurasi secara signifikan [7].

Disisi lain, EfficientNetV2 yang diperkenalkan oleh Google Research dirancang menggunakan *training-aware neural architecture search* dan *progressive learning* untuk mencapai keseimbangan optimal antara efisiensi parameter dan kecepatan pelatihan. Arsitektur EfficientNetV2-B0 menggunakan blok *Fused-MBConv* yang mampu mengurangi latensi dan mempercepat konvergensi tanpa kehilangan performa akurasi [8]. Dalam beberapa studi, EfficientNetV2 terbukti melampaui performa MobileNetV3 pada berbagai dataset [9].

Penelitian ini menyajikan studi komparatif terhadap dua arsitektur *lightweight convolutional neural network*, yaitu MobileNetV3-Large dan EfficientNetV2-B0, dalam konteks pengenalan alfabet BISINDO berbasis representasi citra skeleton dua dimensi hasil ekstraksi landmark MediaPipe.

Dalam literatur pengenalan bahasa isyarat, sebagian studi sebelumnya lebih menekankan evaluasi performa klasifikasi berbasis citra RGB atau data video. Dalam konteks tersebut, penelitian ini menyajikan perbandingan arsitektur CNN ringan menggunakan representasi skeleton dua dimensi, dengan evaluasi yang tidak hanya mencakup performa klasifikasi, tetapi juga aspek stabilitas pelatihan dan pengujian, ketahanan terhadap variasi visual ringan, serta efisiensi komputasi yang diukur melalui jumlah parameter, ukuran model, FLOPs, dan waktu inferensi.

Dengan pendekatan ini, penelitian tidak bertujuan untuk mengusulkan arsitektur baru, melainkan menyajikan analisis komprehensif dan terkontrol terhadap karakteristik kinerja dan efisiensi dua model CNN ringan dalam satu pipeline eksperimen yang konsisten. Hasil analisis diharapkan dapat memberikan referensi metodologis bagi pengembangan sistem pengenalan bahasa isyarat Indonesia yang efisien dan presisi, khususnya untuk aplikasi berbasis kamera dengan kebutuhan pemrosesan mendekati real-time [10].

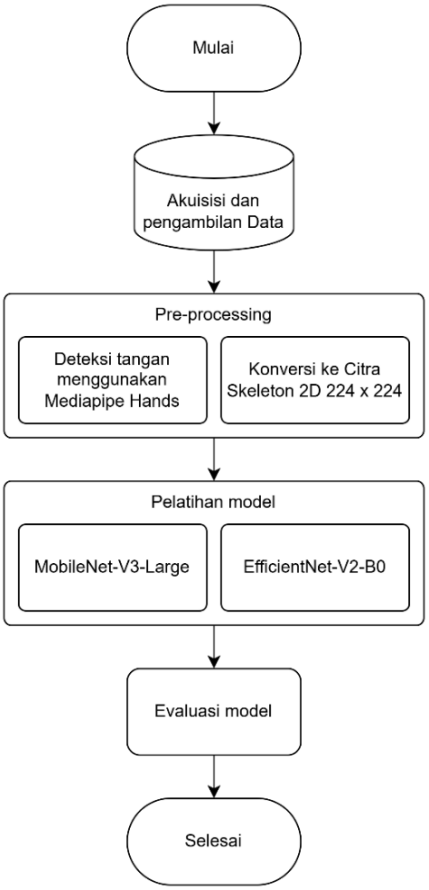
II. METODE

Penelitian ini dirancang untuk melakukan analisis komparatif antara arsitektur MobileNetV3-Large dan EfficientNetV2-B0 dalam pengenalan huruf Bahasa Isyarat Indonesia (BISINDO) berbasis representasi citra skeleton dua dimensi yang dihasilkan dari ekstraksi landmark tangan menggunakan MediaPipe Hands. Pendekatan ini dipilih untuk menekankan pemanfaatan informasi struktural tangan dan mengurangi pengaruh variasi pencahayaan, latar belakang, serta tekstur visual.

Metodologi penelitian disusun sebagai sebuah pipeline terintegrasi yang mencakup tahap persiapan dataset, pra-pemrosesan dan ekstraksi landmark tangan, konversi landmark ke citra 2D, pelatihan model MobileNetV3-Large dan EfficientNetV2-B0, serta evaluasi kinerja dan efisiensi komputasi kedua model. Setiap tahap dirancang untuk memastikan proses pembelajaran yang konsisten dan evaluasi yang adil, sehingga perbandingan performa kedua arsitektur dapat dianalisis secara objektif.

Melalui pipeline ini, penelitian tidak hanya membandingkan performa klasifikasi, tetapi juga mengkaji trade-off antara akurasi dan efisiensi komputasi pada sistem pengenalan huruf BISINDO berbasis landmark, yang relevan untuk pengembangan aplikasi pengenalan bahasa isyarat berbasis kamera secara real-time.

Diagram alur keseluruhan tahapan penelitian ditunjukkan pada Gambar 1.



Gambar 1. Diagram alur penelitian

Keterangan: Diagram ini menggambarkan alur proses penelitian, dimulai dari input citra tangan, ekstraksi landmark MediaPipe, konversi ke citra skeleton 2D, pelatihan model MobileNetV3-Large dan EfficientNetV2-B0, evaluasi dan perbandingan performa.

A. Persiapan Dataset

Dataset yang digunakan dalam penelitian ini berasal dari kumpulan video alfabet Bahasa Isyarat Indonesia (BISINDO) yang diperoleh dari berbagai sumber, termasuk repositori publik Kaggle, platform YouTube, serta rekaman mandiri yang dilakukan oleh peneliti dan partisipan tambahan. Video dari repositori Kaggle terdiri dari dua subjek dengan format video .mov dan .mp4, resolusi sekitar 720p dengan jumlah 2 video, serta kondisi pencahayaan yang relatif cerah, meskipun sebagian video menunjukkan adanya motion blur. Video yang diperoleh dari platform YouTube umumnya memiliki resolusi 1080p dengan kondisi pencahayaan yang relatif cerah dan terdiri dari tiga subjek, berjumlah 3 video. Sementara itu, video hasil perekaman mandiri direkam menggunakan webcam dengan resolusi 1080p tediri dari dua subjek dengan jumlah 3 video, jarak kamera terhadap subjek berkisar antara 75–180 cm, dan kondisi pencahayaan ruangan yang cerah.

Gambar 2. Contoh citra subjek

Seluruh video memiliki frame rate 30 fps dan berdurasi 3 sampai 15 detik pada masing-masing alfabet BISINDO A-Z dan melibatkan total tujuh subjek yang berbeda. Video-video tersebut kemudian diproses menggunakan MediaPipe Hands untuk mengekstraksi 21 titik landmark tangan pada setiap frame. Landmark yang terdeteksi selanjutnya dikonversi menjadi representasi citra skeleton dua dimensi pada latar belakang hitam.

Melalui proses ekstraksi tersebut, diperoleh total 10,309 citra skeleton yang merepresentasikan alfabet BISINDO A–Z. Distribusi jumlah citra per kelas berada pada kisaran yang relatif seimbang, dengan jumlah citra per huruf berkisar antara 317 hingga 469 gambar. Meskipun tidak dilakukan pembalansan data hingga setiap kelas memiliki jumlah yang identik, variasi jumlah citra antar kelas tetap berada dalam rentang yang wajar, sehingga tidak menimbulkan ketimpangan distribusi yang signifikan dalam proses pelatihan dan evaluasi model.

Dataset dibagi ke dalam tiga subset, yaitu data pelatihan, validasi, dan pengujian dengan rasio masing-masing sebesar 70%, 20%, dan 10%. Proses pembagian dilakukan secara acak pada setiap kelas berdasarkan struktur direktori, sehingga distribusi data antar kelas tetap terjaga. Setiap citra ditempatkan ke dalam subset yang sesuai dan diberi label kelas berdasarkan folder asal untuk keperluan pelatihan model klasifikasi multi-kelas.

TABEL I
KONFIGURASI PEMBAGIAN DATASET

Set	Jumlah file	Persentase
Train	7.205	70%
Validation	2.059	20%
Test	1.045	10%

Pembagian dataset dilakukan secara terpisah untuk setiap kelas alfabet BISINDO (A–Z) guna memastikan bahwa distribusi data pada masing-masing subset tetap merepresentasikan seluruh kelas. Dataset hasil pembagian selanjutnya digunakan sebagai masukan pada tahap pelatihan dan evaluasi model, sedangkan proses pengkodean label ke dalam bentuk one-hot encoding dilakukan secara otomatis pada tahap pemuatan data menggunakan utilitas TensorFlow.

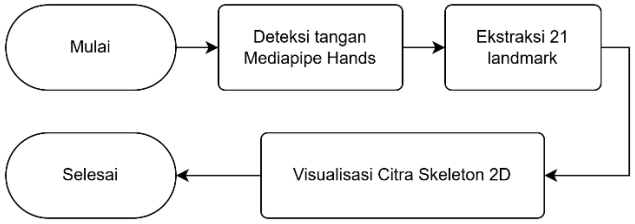
TABEL II
DISTRIBUSI DATASET BISINDO (A-Z)

Parameter	Nilai
Jumlah kelas	26 (A-Z)
Total citra	10.309
Jumlah subjek	7
Rata-rata citra per kelas	317 - 469
Sumber data	Kaggle, YouTube, Rekaman Mandiri
Format data awal	Video (.mov, .mp4)
Frame rate video	30 fps
Resolusi video	720p – 1080p
Resolusi citra hasil preprocessing	224 × 224

B. Pra-Pemrosesan Dan Ekstraksi Landmark

Tahap pra-pemrosesan dilakukan menggunakan MediaPipe Hands untuk mendeteksi 21 titik landmark tangan dari setiap frame video BISINDO. Landmark yang terdeteksi tidak digunakan secara langsung sebagai fitur numerik, melainkan dikonversi menjadi representasi visual berupa citra skeleton dua dimensi. Pendekatan ini memungkinkan model CNN memfokuskan pembelajaran pada pola spasial struktur tangan tanpa dipengaruhi oleh tekstur atau latar belakang [11].

Untuk mengurangi redundansi antar frame yang berurutan, proses ekstraksi diterapkan dengan frame sampling, di mana satu dari setiap lima frame video (frame skip = 5) diproses dan disimpan sebagai citra skeleton. Hasil tahap pra-pemrosesan ini berupa kumpulan citra skeleton dua dimensi yang selanjutnya digunakan pada tahap pelatihan dan evaluasi model klasifikasi.

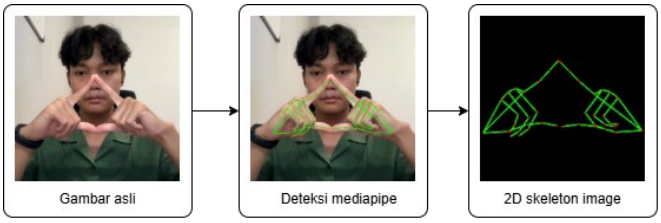


Gambar 3. Alur pra-pemrosesan dan ekstraksi landmark

C. Konversi Landmark ke Citra 2D

Landmark tangan hasil ekstraksi MediaPipe dikonversi menjadi representasi citra skeleton dua dimensi pada latar belakang hitam. Setiap titik landmark dan koneksi antar titik digambarkan mengikuti struktur anatomis tangan, sehingga membentuk skeleton tangan yang merepresentasikan pola gestur. Untuk menjaga konsistensi skala dan posisi, dilakukan proses cropping berbasis bounding box landmark yang dipusatkan pada area tangan. Area hasil cropping kemudian dipadatkan menjadi bentuk persegi dengan penambahan padding yang proporsional, sebelum diubah ukurannya menjadi 224 × 224 piksel sesuai dengan kebutuhan input model CNN.

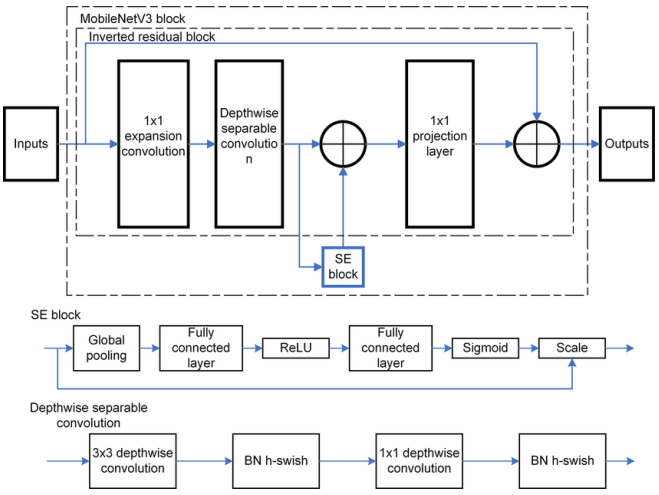
Pendekatan ini memastikan bahwa representasi skeleton tangan terjaga dalam skala dan posisi yang konsisten serta terisi secara optimal pada setiap citra masukan, sehingga variasi ukuran akibat perbedaan jarak kamera atau posisi tangan dapat diminimalkan. Representasi citra skeleton ini selanjutnya digunakan sebagai masukan pada tahap pelatihan model.



Gambar 4. Contoh hasil konversi landmark ke citra 2D (skeleton)

D. Pelatihan Model MobileNetV3-Large

Model MobileNetV3-Large digunakan sebagai salah satu arsitektur lightweight CNN untuk mengenali pola gestur tangan. Model ini menggabungkan blok *depthwise separable convolution*, *Squeeze-and-Excitation (SE)*, dan fungsi aktivasi *h-swish* untuk mencapai efisiensi tinggi tanpa mengorbankan akurasi [12].



Gambar 5. Ilustrasi blok utama MobileNetV3 dengan depthwise separable convolution, SE

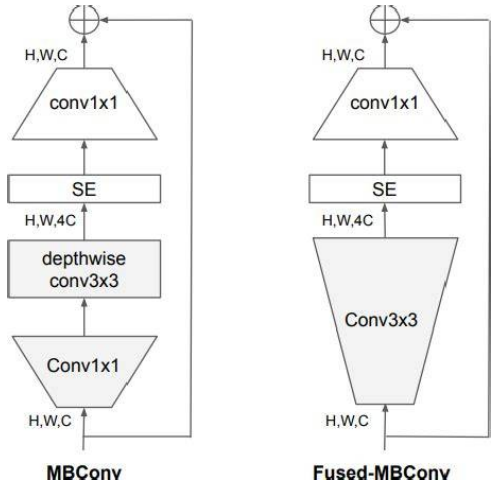
Proses pelatihan model MobileNetV3-Large dilakukan menggunakan pendekatan transfer learning dengan memanfaatkan bobot awal hasil pra-pelatihan pada dataset ImageNet. Seluruh lapisan backbone MobileNetV3-Large dibekukan (frozen) pada tahap pelatihan untuk mempertahankan representasi fitur umum, sementara lapisan klasifikasi dilatih menggunakan dataset BISINDO. Citra masukan diubah ukurannya menjadi 224 × 224 piksel dan diproses menggunakan fungsi preprocessing bawaan MobileNetV3 sebelum diteruskan ke backbone jaringan.

Untuk meningkatkan kemampuan generalisasi model, diterapkan augmentasi data ringan yang mencakup rotasi acak, translasi kecil, dan zoom acak pada tahap pelatihan. Model dilatih menggunakan optimizer Adam dengan learning rate sebesar 0.001, batch size 16, dan maksimum 30 epoch. Mekanisme early stopping berbasis validation loss dengan nilai patience sebesar 10 digunakan untuk mencegah overfitting, serta bobot terbaik model disimpan selama proses pelatihan. Lapisan dropout dengan rasio 0.3 diterapkan sebelum lapisan klasifikasi akhir untuk meningkatkan regularisasi model. Proses ekstraksi fitur dilakukan secara otomatis oleh backbone MobileNetV3-Large yang telah dipra-latih pada dataset ImageNet.

E. Pelatihan Model EfficientNetV2-B0

Model EfficientNetV2-B0 digunakan sebagai pembanding dengan arsitektur MobileNetV3-Large. EfficientNetV2-B0 dikembangkan menggunakan konsep *progressive learning* dan blok *Fused-MBConv* yang meningkatkan efisiensi dan kecepatan pelatihan [13].

Blok Fused-MBConv menggantikan blok MBConv konvensional pada beberapa tahap awal jaringan. Perbedaannya terletak pada penghapusan operasi *depthwise convolution* dan penggabungan operasi *convolution* serta *batch normalization* ke dalam satu langkah, yang mempercepat proses pelatihan tanpa penurunan akurasi. Sementara itu, blok MBConv tetap dipertahankan pada bagian akhir untuk mempertahankan kemampuan ekstraksi fitur mendalam.



Gambar 6. Ilustrasi struktur MBConv dan Fused-MbConv

Pada penelitian ini, EfficientNetV2-B0 diinisialisasi menggunakan bobot pra-latih dari dataset ImageNet dan digunakan sebagai backbone ekstraksi fitur. Seluruh lapisan backbone dibekukan (frozen) selama proses pelatihan, sehingga model beroperasi dalam skema feature extraction. Lapisan klasifikasi tambahan dilatih menggunakan dataset BISINDO. Citra masukan diubah ukurannya menjadi 224 × 224 piksel dan diproses menggunakan fungsi preprocessing bawaan EfficientNetV2 sebelum diteruskan ke backbone jaringan.

Untuk meningkatkan kemampuan generalisasi, diterapkan augmentasi data ringan berupa rotasi acak, translasi kecil, dan zoom acak pada tahap pelatihan. Model dilatih menggunakan optimizer Adam dengan learning rate sebesar 0.001, batch size 16, dan maksimum 30 epoch. Mekanisme early stopping berbasis validation loss dengan nilai patience sebesar 10 diterapkan untuk mencegah overfitting, serta lapisan dropout dengan rasio 0.3 digunakan sebelum lapisan klasifikasi akhir.

Pendekatan pelatihan ini dibuat identik dengan konfigurasi MobileNetV3-Large guna memastikan perbandingan performa yang adil antara kedua arsitektur, baik dari sisi akurasi klasifikasi maupun efisiensi komputasi.

F. Evaluasi Dan Perbandingan Model

Evaluasi performa model dilakukan menggunakan dataset uji (test set) yang mencakup 10% dari total data, yang tidak digunakan selama proses pelatihan maupun validasi. Penggunaan data uji bertujuan untuk mengukur kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Pengukuran performa dilakukan menggunakan empat metrik utama, yaitu Accuracy, Precision, Recall, dan F1-Score [14]. Setiap metrik dihitung berdasarkan hasil prediksi model pada data uji.

Metode ini, menganalisis performa model berdasarkan empat kategori berikut.

1) True Positive (TP): data positif diprediksi benar.
2) True Negatif (TN): data negatif diprediksi benar.
3) False Positive (FP): data negatif salah diprediksi sebagai positif.
4) False Negatif (FN): data positif salah diprediksi sebagai negative.

Empat kategori tersebut akan digunakan sebagai dasar untuk menghitung performa, seperti berikut:

1) Akurasi: Tingkat keakuratan dari sebuah model dalam mengklasifikasi data, dapat dihitung berdasarkan rumus berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2) Presisi: Tingkat ketepatan model dalam memprediksi kelas yang diminta, berikut rumus dari presisi.

$$Precision = \frac{TP}{TP + FP}$$

3) Recall: Tingkat sensitivitas yang menunjukkan seberapa baik model dalam menemukan kembali informasi yang benar dari dataset, berikut rumus dari recall.

$$Recall = \frac{TP}{TP + FN}$$

4) F1-Score: Tingkat kesimbangan antara presisi dan recall, ditunjukkan dengan rumus berikut.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Selain evaluasi berbasis metrik klasifikasi konvensional, penelitian ini menerapkan *metamorphic testing* untuk mengevaluasi ketahanan model terhadap variasi visual ringan pada citra skeleton dua dimensi. Pendekatan ini digunakan untuk menilai konsistensi prediksi model ketika citra masukan mengalami transformasi visual tertentu tanpa memerlukan label kebenaran tambahan.

Empat jenis relasi metamorfik digunakan dalam penelitian ini, yaitu rotasi, translasi, perubahan skala terbatas, dan penambahan noise Gaussian ringan. Transformasi diterapkan secara langsung pada citra skeleton dua dimensi yang digunakan sebagai masukan model klasifikasi CNN.

Perubahan skala dibatasi karena citra skeleton telah dinormalisasi melalui cropping berbasis landmark dan padding proporsional pada tahap pra-pemrosesan. Seluruh pengujian metamorphic dilakukan menggunakan dataset uji yang sama dengan evaluasi klasifikasi akhir, tanpa melibatkan proses pelatihan ulang model. Konsistensi prediksi diukur menggunakan consistency percentage, yaitu persentase jumlah sampel yang menghasilkan label prediksi yang sama sebelum dan sesudah transformasi diterapkan terhadap total jumlah sampel uji.

Untuk memastikan bahwa perbedaan performa antara MobileNetV3-Large dan EfficientNetV2-B0 tidak terjadi secara kebetulan, penelitian ini menerapkan uji signifikansi statistik menggunakan metode *paired bootstrap confidence interval* pada dataset uji. Evaluasi dilakukan dengan skema *resampling* berpasangan sebanyak 10.000 iterasi terhadap prediksi kedua model pada sampel uji yang sama, sehingga distribusi empiris selisih performa dapat diestimasi secara non-parametrik.

Pendekatan *paired bootstrap* dipilih karena mampu mempertahankan keterkaitan antar prediksi model pada setiap sampel uji serta tidak bergantung pada asumsi distribusi tertentu. Interval kepercayaan 95% dihitung dari distribusi bootstrap tersebut dan digunakan sebagai dasar untuk menilai signifikansi perbedaan performa antara kedua arsitektur berdasarkan metrik *Accuracy* dan *F1-score*.

III. HASIL DAN PEMBAHASAN

A. Hasil Pelatihan Model

Setelah proses pelatihan kedua model, MobileNetV3-Large mencatat training accuracy sebesar 93.25% dan validation accuracy sebesar 90.48%, sedangkan

EfficientNetV2-B0 mencapai training accuracy sebesar 93.83% dengan validation accuracy sebesar 95.58%.

Berdasarkan hasil evaluasi pada dataset uji (test set), MobileNetV3-Large menghasilkan accuracy sebesar 89.67%. EfficientNetV2-B0 memperoleh accuracy sebesar 95.98%, Sementara itu,

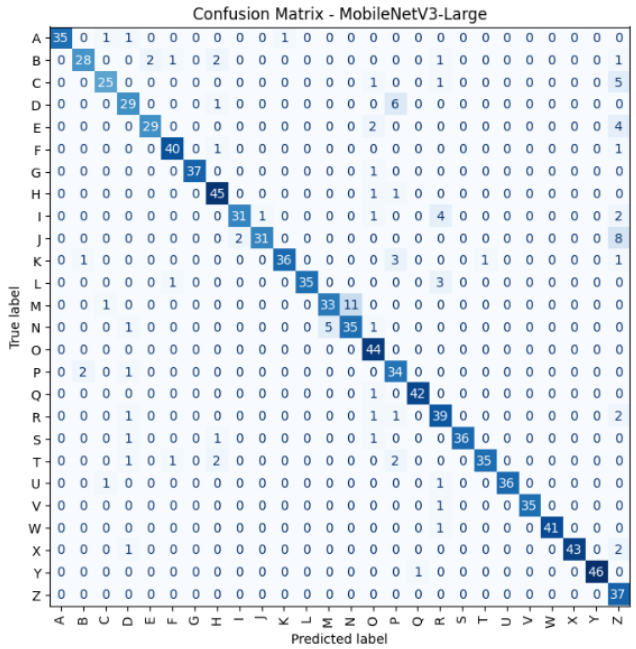
TABEL III
RINGKASAN HASIL PELATIHAN MODEL

Model	Train Acc	Val Acc	Test Acc
MobileNetV3-Large	93.25%	90.48%	89.67%
EfficientNetV2-B0	93.83%	95.58%	95.98%

B. Analisis Kinerja Model

Evaluasi kinerja model dilakukan menggunakan *confusion matrix* untuk menganalisis performa klasifikasi pada tingkat per kelas serta mengidentifikasi pola kesalahan yang terjadi pada kedua model. Matriks ini merepresentasikan jumlah prediksi benar dan salah untuk setiap kombinasi antara kelas aktual dan kelas hasil prediksi pada dataset uji.

Berdasarkan confusion matrix yang diperoleh, kedua model menunjukkan dominasi nilai pada diagonal utama, yang mengindikasikan tingkat prediksi benar yang tinggi pada sebagian besar kelas alfabet BISINDO. Meskipun demikian, masih terdapat sejumlah misklasifikasi pada kelas-kelas tertentu yang memiliki kemiripan konfigurasi jari.



Gambar 7. Confusion matrix MobileNetV3-Large

Pada model MobileNetV3-Large, confusion matrix menunjukkan konsentrasi nilai yang tinggi pada diagonal utama untuk sebagian besar kelas alfabet BISINDO, yang menandakan dominasi jumlah prediksi benar pada mayoritas kelas. Meskipun demikian, masih ditemukan sejumlah kesalahan klasifikasi pada beberapa pasangan kelas tertentu.

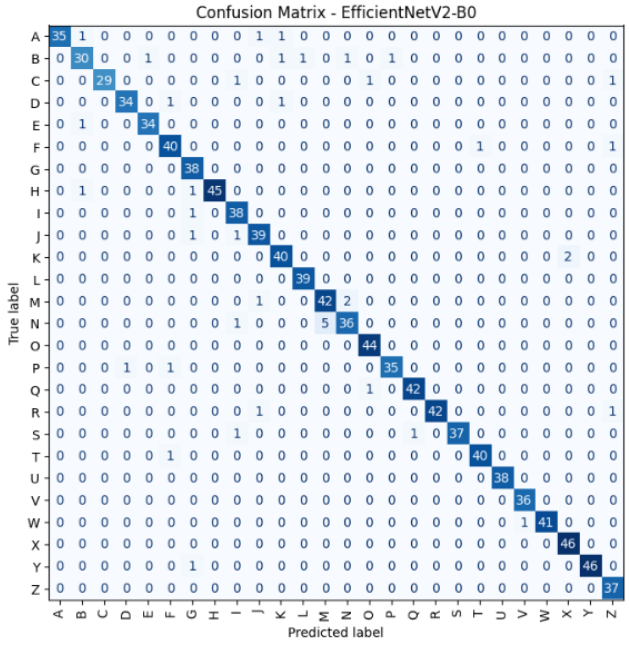
Kesalahan paling dominan terjadi pada kelas M yang diprediksi sebagai kelas N sebanyak 11 sampel, serta kelas J yang diprediksi sebagai kelas Z sebanyak 8 sampel. Selain itu, kesalahan dengan frekuensi relatif tinggi juga ditemukan pada kelas D → P sebanyak 6 sampel, serta C → Z dan N → M yang masing-masing tercatat sebanyak 5 sampel. Kesalahan tambahan dengan frekuensi yang lebih rendah tercatat pada beberapa pasangan kelas lain, seperti E → Z dan I → R (masing-masing 4 sampel), serta L → R dan K → P (masing-masing 3 sampel), dan P → B sebanyak 2 sampel.

Ringkasan kelas dengan tingkat kesalahan tertinggi pada model MobileNetV3-Large disajikan pada Tabel IV.

TABEL IV
TOP 5 KELAS ERROR MOBILE NET V3-LARGE

Kelas	Precision	Recall	F1-Score	Error Rate
M	0.86	0.73	0.79	0.26
J	0.96	0.75	0.84	0.24
C	0.89	0.78	0.83	0.21
I	0.93	0.79	0.86	0.20
B	0.90	0.80	0.84	0.20

Pada model EfficientNetV2-B0, distribusi nilai pada confusion matrix menunjukkan dominasi yang sangat kuat pada diagonal utama, yang menandakan tingkat prediksi benar yang tinggi pada hampir seluruh kelas alfabet BISINDO. Secara umum, jumlah kesalahan klasifikasi yang terjadi pada model ini lebih sedikit dan lebih tersebar dibandingkan MobileNetV3-Large.



Gambar 8. Confusion matrix EfficientNetV2-B0

Kesalahan klasifikasi yang paling menonjol terjadi pada kelas N yang diprediksi sebagai kelas M sebanyak 5 sampel. Selain itu, kesalahan dua arah antara kelas M dan N masih ditemukan, namun dengan frekuensi yang lebih rendah, yaitu M → N sebanyak 2 sampel. Kesalahan dengan frekuensi

relatif rendah juga terjadi pada kelas K → X sebanyak 2 sampel. Kesalahan lain dengan frekuensi sangat rendah (masing-masing 1 sampel) meliputi pasangan A → B, A → J, J → I, M → J, N → I, serta P → D dan P → F.

Ringkasan kelas dengan tingkat kesalahan tertinggi pada model EfficientNetV2-B0 disajikan pada Tabel V.

TABEL V
TOP 5 KELAS ERROR EFFICIENT NET-V2B0

Kelas	Precision	Recall	F1-Score	Error Rate
N	0.92	0.85	0.88	0.14
B	0.90	0.85	0.88	0.14
C	1.00	0.90	0.95	0.09
A	1.00	0.92	0.95	0.07
M	0.89	0.93	0.91	0.06

Sebagai ringkasan performa klasifikasi pada tingkat global, evaluasi dilakukan pada dataset uji menggunakan metrik accuracy, recall, dan F1-score. Hasil evaluasi global kedua model dirangkum dalam Tabel VI.

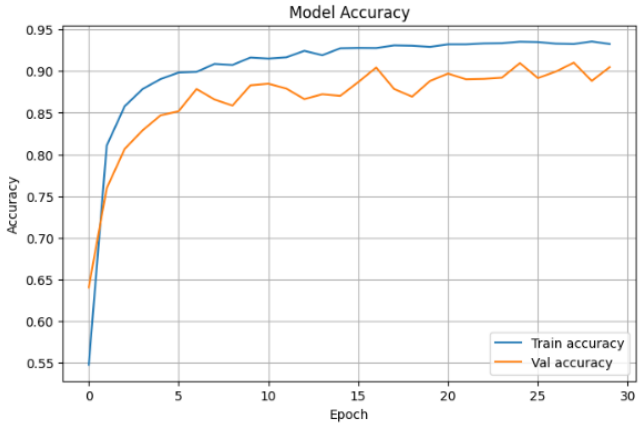
TABEL VI
HASIL EVALUASI PADA DATASET UJI

Model	Accuracy	Recall	F1-Score
MobileNetV3-Large	89.67%	89.46%	89.76%
EfficientNetV2-B0	95.98%	95.90%	95.93%

Berdasarkan hasil pada Tabel VI, MobileNetV3-Large memperoleh nilai accuracy 89.67% dan recall sebesar 89,46% dengan F1-score sebesar 89,76% pada dataset uji, sedangkan EfficientNetV2-B0 mencapai nilai accuracy 95.98%, recall 95.90%, dan F1-score sebesar 95,93%.

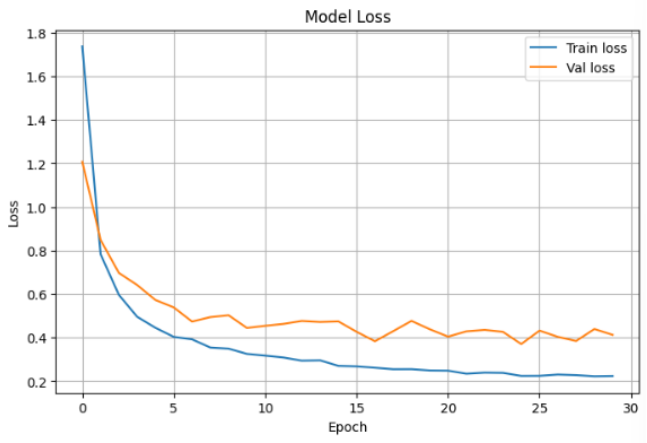
C. *Perbandingan Arsitektur dan Evaluasi*

Untuk memahami dinamika pelatihan, dilakukan analisis perbandingan kurva training–validation accuracy dan training–validation loss dari kedua model.



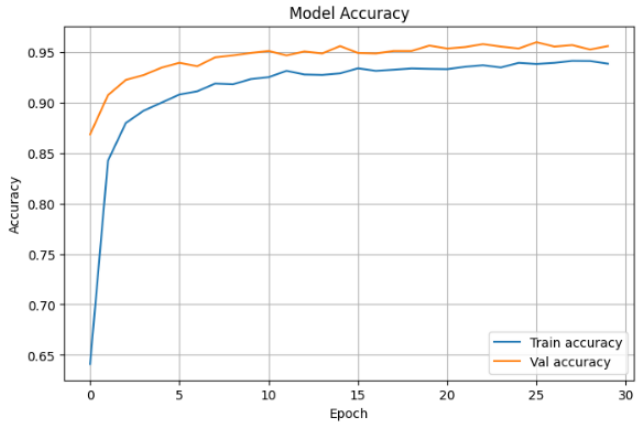
Gambar 9. Grafik training vs validation accuracy MobileNetV3-Large

Pada model MobileNetV3-Large, kurva *training accuracy* menunjukkan peningkatan yang tajam pada beberapa epoch awal, kemudian meningkat secara bertahap hingga mencapai nilai sekitar 93–94% pada epoch akhir. Kurva *validation accuracy* juga mengalami tren peningkatan yang stabil dan berada pada kisaran 89–91% menjelang akhir proses pelatihan.



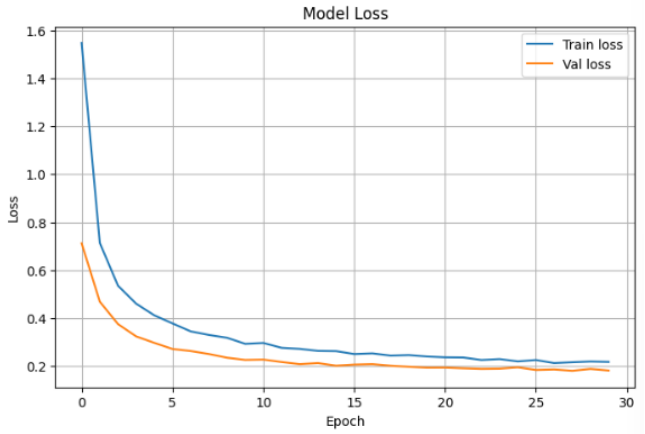
Gambar 10. Grafik training vs validation loss MobileNetV3-Large

Grafik *training loss* memperlihatkan penurunan yang konsisten dari nilai awal yang tinggi hingga mendekati 0.22 pada epoch akhir, sementara *validation loss* turut menurun dan berada pada kisaran 0.38–0.45 dengan fluktuasi yang relatif kecil sepanjang proses pelatihan. Pola kurva *accuracy* dan *loss* untuk model MobileNetV3-Large ditampilkan pada Gambar 9 dan Gambar 10.



Gambar 11. Grafik training vs validation accuracy EfficientNetV2-B0

Pada model EfficientNetV2-B0, kurva *training accuracy* dan *validation accuracy* menunjukkan peningkatan yang cepat pada beberapa epoch awal, kemudian mencapai kondisi relatif stabil pada epoch-epoch selanjutnya. Nilai *validation accuracy* berada pada kisaran 95–96% pada akhir proses pelatihan. Kurva *training loss* dan *validation loss* juga memperlihatkan penurunan yang konsisten sepanjang proses pelatihan hingga mencapai kisaran 0.18–0.21 pada epoch akhir. Pola kurva *accuracy* dan *loss* untuk model EfficientNetV2-B0 ditampilkan pada Gambar 11 dan Gambar 12.



Gambar 12. Grafik training vs validation loss EfficientNetV2-B0

Perbandingan kurva pelatihan kedua model menunjukkan adanya perbedaan nilai akhir *accuracy* dan *loss* pada data validasi, yang mencerminkan variasi performa pelatihan antara MobileNetV3-Large dan EfficientNetV2-B0 dalam konfigurasi eksperimen yang digunakan.

TABEL VII
RINGKASAN KOMPLEKSITAS MODEL

Model	Parameter	Model Size	FLOPs	Inference Time
MobileNet V3-Large	3.02 M	12.4 MB	0.44 GFLOPs	0.0343
EfficientNet V2-B0	5.95 M	23.9 MB	1.46 GFLOPs	0.0381

Tabel VII menyajikan ringkasan kompleksitas model dan performa inferensi dari MobileNetV3-Large dan EfficientNetV2-B0. MobileNetV3-Large memiliki jumlah parameter sebesar 3.021.338 dengan ukuran model 12,4 MB serta kompleksitas komputasi 443.594.228 FLOPs. Model ini mencatat waktu inferensi rata-rata sebesar 0,0343 detik per citra, Inference time diukur sebagai rata-rata waktu forward pass per citra (batch size = 1) pada data uji

Sebaliknya, EfficientNetV2-B0 memiliki kompleksitas yang lebih tinggi dengan 5.952.618 parameter, ukuran model 23,9 MB, dan kebutuhan komputasi mencapai 1.455.382.223 FLOPs. Waktu inferensi yang dihasilkan sedikit lebih tinggi, yaitu 0,0381 detik per citra.

D. Hasil Uji Signifikansi Statistik (Bootstrap CI)

Untuk mengevaluasi apakah perbedaan performa antara MobileNetV3-Large dan EfficientNetV2-B0 bersifat signifikan secara statistik, dilakukan uji signifikansi menggunakan metode *paired bootstrap confidence interval* pada dataset uji. Pengujian dilakukan dengan skema resampling berpasangan sebanyak 10.000 iterasi terhadap prediksi kedua model pada sampel uji yang sama. Evaluasi difokuskan pada dua metrik utama, yaitu Accuracy dan F1-score (macro-average).

TABEL VIII
BOOTSTRAP CI: AKURASI DAN F1

Metrik	Δ Mean (EF−MN)	95% CI (Low)	95% CI (Up)	p-value
Accuracy	0,0632	0,0469	0,0794	< 0,0001
F1-macro	0,0623	0,0461	0,0789	< 0,0001

Keterangan: Δ menunjukkan selisih performa antara EfficientNetV2-B0 dan MobileNetV3-Large (Δ = EfficientNetV2-B0 − MobileNetV3-Large). Interval kepercayaan 95% diperoleh melalui metode *paired bootstrap resampling* sebanyak 10.000 iterasi pada dataset uji. Nilai p dilaporkan sebagai batas bawah resolusi numerik berdasarkan jumlah iterasi bootstrap.

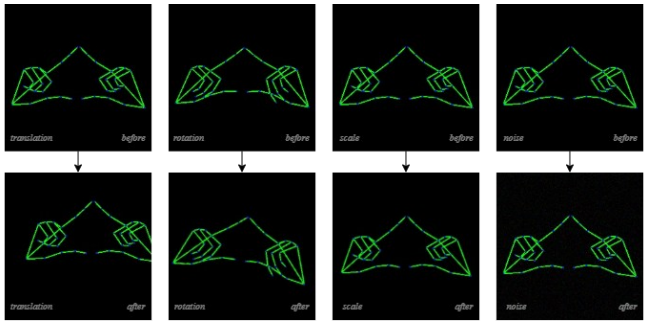
Berdasarkan hasil uji *paired bootstrap confidence interval* yang disajikan pada Tabel VIII, EfficientNetV2-B0 menunjukkan performa yang lebih tinggi dibandingkan MobileNetV3-Large pada kedua metrik evaluasi. Pada metrik Accuracy, selisih rata-rata performa antara kedua model sebesar 6,32%, dengan interval kepercayaan 95% berada pada rentang [4,69%, 7,94%]. Seluruh rentang interval kepercayaan berada di atas nol, yang mengindikasikan bahwa perbedaan performa tersebut signifikan secara statistik ($p < 0,0001$).

Hasil yang konsisten juga diperoleh pada metrik F1-score (macro-average), dengan selisih rata-rata sebesar 6,23% dan interval kepercayaan 95% pada rentang [4,61%, 7,89%]. Interval kepercayaan yang sepenuhnya berada di atas nol menunjukkan bahwa perbedaan performa antara kedua model bersifat konsisten dan tidak disebabkan oleh variasi acak pada data uji.

E. Hasil Metamorphic Testing

Metamorphic testing diterapkan untuk mengevaluasi konsistensi prediksi model klasifikasi CNN terhadap variasi transformasi visual ringan pada citra skeleton dua dimensi.

Untuk setiap relasi metamorfik digunakan dua tingkat transformasi, yaitu 5% dan 10%. Konsistensi prediksi diukur sebagai persentase jumlah sampel yang menghasilkan label prediksi yang sama sebelum dan sesudah transformasi diterapkan.



Gambar 13. Contoh gambar penambahan transformasi untuk Metamorphic test

TABEL IX
HASIL METAMORPHIC TESTING PADA DATA UJI

Model	Relation	Lvl (%)	Correct	Cons. (%)
MobileNet V3-Large	Rot.	5 / 10	977 / 964	93.4 / 92.2
	Trans.	5 / 10	966 / 951	92.4 / 91.0
	Scale	5 / 10	962 / 971	92.0 / 92.9
	Noise	5 / 10	946 / 913	90.5 / 87.3
EfficientNet V2-B0	Rot.	5 / 10	1023 / 1020	97.8 / 97.6
	Trans.	5 / 10	1013 / 1007	96.9 / 96.3
	Scale	5 / 10	1022 / 1015	97.8 / 97.1
	Noise	5 / 10	962 / 960	92.0 / 91.8

Keterangan: Rot. = Rotation, Trans. = Translation, Lvl = tingkat transformasi (%), Correct = jumlah prediksi, Cons. = persentase konsistensi prediksi. Nilai ditampilkan untuk dua tingkat transformasi (5% / 10%). Pengujian metamorphic testing dilakukan pada subset data uji yang terdiri dari 1.045 citra skeleton, yang sama dengan data uji pada evaluasi klasifikasi akhir.

Tabel IX menyajikan hasil *metamorphic testing* pada data uji BISINDO dengan menerapkan empat jenis relasi metamorfik, yaitu rotasi, translasi, perubahan skala, dan penambahan *Gaussian noise*, pada dua tingkat transformasi (5% dan 10%). Untuk setiap relasi, dicatat jumlah prediksi yang konsisten serta persentase konsistensi prediksi pada masing-masing model.

Secara umum, konsistensi prediksi pada tingkat transformasi 10% cenderung lebih rendah dibandingkan dengan tingkat 5% pada sebagian besar relasi metamorfik yang diuji. Meskipun demikian, kedua model masih mempertahankan tingkat konsistensi yang relatif tinggi pada seluruh jenis transformasi. Nilai konsistensi prediksi untuk setiap relasi dan tingkat transformasi ditampilkan secara rinci pada Tabel IX.

F. Interpretasi Hasil

Berdasarkan hasil evaluasi pada dataset uji, EfficientNetV2-B0 secara konsisten mencapai nilai *accuracy*, *recall*, dan *F1-score* yang lebih tinggi dibandingkan MobileNetV3-Large. Perbedaan performa ini menunjukkan bahwa EfficientNetV2-B0 memiliki kemampuan generalisasi yang lebih baik dalam mengenali pola gestur alfabet BISINDO berbasis representasi skeleton dua dimensi, meskipun kedua model dilatih menggunakan konfigurasi yang identik.

Analisis *confusion matrix* menunjukkan bahwa kesalahan klasifikasi pada kedua model terkonsentrasi pada pasangan kelas dengan kemiripan konfigurasi jari yang tinggi, khususnya M–N. Pola kesalahan dua arah ini mengindikasikan bahwa perbedaan gestur yang bersifat struktural halus masih sulit dibedakan dalam representasi skeleton dua dimensi. Dibandingkan MobileNetV3-Large, EfficientNetV2-B0 menunjukkan distribusi kesalahan yang lebih jarang dan tersebar, yang mencerminkan kemampuan diskriminasi fitur yang lebih baik.

Dari sisi kompleksitas, MobileNetV3-Large menawarkan efisiensi komputasi yang lebih tinggi dengan jumlah parameter dan waktu inferensi yang lebih rendah, sedangkan EfficientNetV2-B0 menunjukkan peningkatan kompleksitas yang diiringi oleh peningkatan performa klasifikasi. sejalan dengan temuan pada studi sebelumnya yang menekankan stabilitas pendekatan berbasis skeleton dalam pengenalan gestur tangan [15].

Hasil uji paired bootstrap confidence interval mengindikasikan bahwa perbedaan performa antara MobileNetV3-Large dan EfficientNetV2-B0 bersifat konsisten dan signifikan secara statistik pada dataset uji. Seluruh interval kepercayaan selisih performa yang berada di atas nol menunjukkan bahwa peningkatan nilai Accuracy dan F1-score (macro-average) yang diperoleh EfficientNetV2-B0 tidak disebabkan oleh variasi acak pada data uji, melainkan mencerminkan perbedaan kinerja yang stabil antar model.

Hasil metamorphic testing menunjukkan bahwa kedua model mengalami penurunan konsistensi prediksi seiring meningkatnya tingkat transformasi visual. Namun, EfficientNetV2-B0 mempertahankan tingkat konsistensi yang lebih tinggi dan relatif stabil pada seluruh relasi metamorfik, sedangkan MobileNetV3-Large lebih sensitif terhadap gangguan visual acak, khususnya noise Gaussian. Temuan ini melengkapi evaluasi klasifikasi konvensional dengan menyoroti perbedaan ketahanan prediksi antar arsitektur.

G. Keterbatasan Penelitian

Penelitian ini memiliki sejumlah keterbatasan yang perlu diperhatikan dalam menafsirkan hasil yang diperoleh. Pendekatan pengenalan huruf BISINDO yang digunakan sangat bergantung pada kualitas landmark tangan yang dihasilkan oleh MediaPipe Hands. Pu et al. melaporkan bahwa lebih dari 50% tangan gagal diidentifikasi oleh MediaPipe Hands ketika diagonal motion blur diterapkan, serta menunjukkan bahwa occlusion dan illumination variation merupakan hambatan utama bagi sistem hand pose estimation berbasis visi computer [16].

Evaluasi performa model dilakukan menggunakan dataset dengan jumlah subjek yang terbatas, sehingga hasil yang diperoleh merepresentasikan performa rata-rata pada kumpulan pengguna tersebut. Analisis performa berbasis subjek (*subject-wise evaluation*) belum dilakukan, sehingga variasi performa antar individu belum dapat dikaji secara mendalam, dan ketahanan model terhadap variasi pengguna (inter-user robustness) belum dianalisis secara eksplisit.

Meskipun pengukuran waktu inferensi menunjukkan bahwa kedua model mendukung pemrosesan mendekati *real-time*, penelitian ini belum mengevaluasi sistem secara *end-to-end* pada skenario pengenalan berbasis video *real-time*. Aspek-aspek praktis seperti latensi kumulatif pada tahap ekstraksi landmark, kualitas perangkat akuisisi video, serta kestabilan input pada lingkungan yang tidak terkontrol masih menjadi tantangan dalam penerapan sistem secara nyata.

Ruang lingkup penelitian ini dibatasi pada pengenalan alfabet BISINDO statis dan belum mencakup gestur dinamis, transisi antar tanda, maupun komponen non-manual. Pembatasan ruang lingkup ini dilakukan untuk memfokuskan kajian pada evaluasi arsitektur CNN ringan sebagai fondasi

awal pengembangan sistem pengenalan BISINDO berbasis visi komputer.

IV. KESIMPULAN

Penelitian ini menyajikan analisis komparatif antara dua arsitektur *lightweight convolutional neural network*, yaitu MobileNetV3-Large dan EfficientNetV2-B0, untuk pengenalan huruf Bahasa Isyarat Indonesia (BISINDO) berbasis citra skeleton dua dimensi hasil ekstraksi MediaPipe Hands. Evaluasi difokuskan pada performa klasifikasi, efisiensi komputasi, serta ketahanan model terhadap variasi visual ringan.

Hasil eksperimen menunjukkan bahwa EfficientNetV2-B0 mencapai performa klasifikasi yang lebih tinggi, sedangkan MobileNetV3-Large menawarkan efisiensi komputasi dan kecepatan inferensi yang lebih baik. Temuan ini menegaskan adanya *trade-off* antara akurasi dan efisiensi dalam pemilihan arsitektur CNN ringan untuk pengenalan BISINDO berbasis visi komputer.

Dari perspektif kebutuhan pengguna BISINDO, hasil ini menunjukkan pentingnya pemilihan arsitektur yang seimbang antara akurasi dan efisiensi untuk mendukung sistem berbasis kamera dengan keterbatasan sumber daya. Dalam konteks aplikasi nyata, temuan penelitian ini berpotensi menjadi landasan awal bagi pengembangan sistem asistif, seperti aplikasi pembelajaran alfabet BISINDO atau alat bantu komunikasi dasar berbasis kamera.

Penelitian ini belum mengevaluasi sistem secara *end-to-end* pada skenario video *real-time*. Oleh karena itu, pengembangan sistem BISINDO dinamis, integrasi komponen non-manual, serta optimasi model melalui teknik *quantization* dan *pruning* menjadi arah penelitian selanjutnya.

DAFTAR PUSTAKA

[1] R. Fahlevi and C. Rozikin, "Identifikasi isyarat tangan bisindo dengan algoritma cnn dan transfer learning menggunakan mobilenetv2," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 4, pp. 6592–6597, May 2025, doi: 10.36040/JATI.V9I4.14095.

[2] A. Saleh, "A Comparative Analysis of CNN and SVM for Static Sign Language Recognition Using MediaPipe Landmarks," *Journal of Intelligent System and Telecommunication*, vol. 1, no. 2, pp. 225–238, Jun. 2025, doi: 10.26740/JISTEL.V1N2.P225-238.

[3] R. Sutjiadi, "Android-Based Application for Real-Time Indonesian Sign Language Recognition Using Convolutional Neural Network," *TEM Journal*, vol. 12, no. 3, pp. 1541–1549, Aug. 2023, doi: 10.18421/TEM123-35.

[4] B. Sundar and T. Bagyammal, "American Sign Language Recognition for Alphabets Using MediaPipe and LSTM," *Procedia Comput Sci*, vol. 215, pp. 642–651, Jan. 2022, doi: 10.1016/J.PROCS.2022.12.066.

[5] J. Bora, S. Dehingia, A. Boruah, A. A. Chetia, and D. Gogoi, "Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning," *Procedia Comput Sci*, vol. 218, pp. 1384–1393, Jan. 2023, doi: 10.1016/J.PROCS.2023.01.117.

[6] O. Yusuf, M. Habib, and M. Moustafa, "Real-Time Hand Gesture Recognition: Integrating Skeleton-Based Data Fusion and Multi-Stream CNN," Oct. 2024, Accessed: Oct. 26, 2025. [Online]. Available: https://arxiv.org/pdf/2406.15003v2

[7] D. Joan, V. Vincent, K. J. Daniel, S. Achmad, and R. Sutoyo, "BISINDO Hand-Sign Detection Using Transfer Learning," *8th International Conference on Recent Advances and Innovations in Engineering: Empowering Computing, Analytics, and Engineering Through Digital Innovation, ICRAIE 2023*, pp. 1–7, Dec. 2023, doi: 10.1109/ICRAIE59459.2023.10468194.

[8] T. Shahriar, "Comparative Analysis of Lightweight Deep Learning Models for Memory-Constrained Devices," May 2025, Accessed: Oct. 26, 2025. [Online]. Available: https://arxiv.org/pdf/2505.03303

[9] M. Al-Hammadi *et al.*, "Deep Learning-Based Approach for Sign Language Gesture Recognition With Efficient Hand Gesture Representation," *IEEE Access*, vol. 8, pp. 192527–192542, Oct. 2020, doi: 10.1109/ACCESS.2020.3032140.

[10] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," *Proc Mach Learn Res*, vol. 139, pp. 10096–10106, Apr. 2021, Accessed: Oct. 26, 2025. [Online]. Available: https://arxiv.org/pdf/2104.00298

[11] J. Shin, A. S. M. Miah, M. H. Kabir, M. A. Rahim, and A. Al Shiam, "A Methodological and Structural Review of Hand Gesture Recognition Across Diverse Data Modalities," *IEEE Access*, vol. 12, pp. 142606–142639, 2024, doi: 10.1109/ACCESS.2024.3456436.

[12] S. Sharma and S. Singh, "ISL recognition system using integrated mobile-net and transfer learning method," *Expert Syst Appl*, vol. 221, pp. 119772–119772, Mar. 2023, doi: 10.1016/J.ESWA.2023.119772.

[13] I. Rizka Fadhillah, M. Muharrom Al Haromainy, and H. Maulana, "Implementasi model transfer learning efficientnet untuk pendeteksian bahasa isyarat indonesia (bisindo) pada perangkat android," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 4, pp. 7816–7822, Aug. 2024, doi: 10.36040/JATI.V8I4.10463.

[14] R. A. Lashaki, Z. Raeisi, N. Razavi, M. Goodarzi, and H. Najafzadeh, "Optimized classification of dental implants using convolutional neural networks and pre-trained models with preprocessed data," *BMC Oral Health*, vol. 25, no. 1, pp. 1–22, Dec. 2025, doi: 10.1186/S12903-025-05704-0/TABLES/3.

[15] M. K. Habib, O. Yusuf, and M. Moustafa, "Skeleton-Based Real-Time Hand Gesture Recognition Using Data Fusion and Ensemble Multi-Stream CNN Architecture," *Technologies 2025, Vol. 13, Page 484*, vol. 13, no. 11, p. 484, Oct. 2025, doi: 10.3390/TECHNOLOGIES13110484.

[16] M. Pu, C. Y. Chong, and M. K. Lim, "Robustness Evaluation in Hand Pose Estimation Models using Metamorphic Testing," Mar. 2023, Accessed: Jan. 09, 2026. [Online]. Available: http://arxiv.org/abs/2303.04566