Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3410~3418
e-ISSN: 2548-6861

L2IC and MobileViT-XXS for BISINDO Alphabet Recognition

Chanan Artamma [1]*, Majid Rahardi [2]*
* Informatika, Universitas Amikom Yogyakarta
chanan69@students.amikom.ac.id [1], majid@amikom.ac.id [2]

Article Info

Article history:

Received 2025-10-24
Revised 2025-11-20
Accepted 2025-11-22

Keyword:

Indonesian Sign Language (BISINDO),
Landmark-to-Image Conversion (L2IC),
Deep Learning

ABSTRACT

This study proposes a Landmark-to-Image Conversion (L2IC) approach integrated with the MobileViT-XXS architecture for Indonesian Sign Language (BISINDO) alphabet recognition. The method converts 42 hand keypoints, extracted using MediaPipe Hands into normalized 224×224 grayscale images to capture spatial hand patterns more effectively. These L2IC representations are then used as input to the MobileViT-XXS model, trained for 30 epochs with a learning rate of 0.001. Experimental results show that the model achieves an accuracy and Macro F1-Score of 97.98%, outperforming baseline approaches using raw RGB images and MLP-based classification on numerical keypoints. While the model demonstrates strong performance in controlled offline experiments, further evaluation is required to assess its robustness under real-world dynamic BISINDO usage and deployment on resource-limited devices. These findings indicate that the L2IC representation effectively captures essential spatial information, contributing to high recognition accuracy in static BISINDO hand gesture classification.

I. PENDAHULUAN

Komunikasi merupakan kebutuhan dasar manusia yang memungkinkan interaksi, pertukaran informasi, dan integrasi social. Namun, tidak semua individu memiliki kemampuan komunikasi yang setara. Berdasarkan hasil Sensus Penduduk Long Form 2020 yang dirilis oleh Badan Pusat Statistik (BPS), sekitar 1,43% penduduk Indonesia merupakan penyandang disabilitas, dengan 0,36% mengalami gangguan pendengaran dan 0,35% mengalami gangguan bicara [1]. Kondisi ini menunjukkan bahwa kelompok tunarungu dan tunawicara menghadapi hambatan komunikasi yang signifikan dalam kehidupan sehari-hari. Menurut laporan World Health Organization (WHO), lebih dari 1,5 miliar orang di seluruh dunia menderita gangguan pendengaran, dan jumlah ini terus meningkat seiring dengan penuaan populasi [2]. Hambatan komunikasi tersebut tidak hanya berdampak pada keterbatasan akses pendidikan dan pekerjaan, tetapi juga menghambat partisipasi sosial dan integrasi penuh dalam masyarakat [3].

Untuk mengatasi hambatan tersebut, Bahasa Isyarat Indonesia (BISINDO) berperan penting sebagai sarana komunikasi utama bagi penyandang tunarungu di Indonesia. Namun, keterbatasan jumlah penerjemah professional menyebabkan komunikasi antara penyandang disabilitas dengan masyarakat umum masih sulit dilakukan [4]. Kondisi ini mendorong berbagai penelitian di bidang kecerdasan buatan (Artificial Intelligence) dan visi komputer (Computer Vision) untuk mengembangkan sistem penerjemah otomatis yang mampu mengenali bahasa isyarat secara real-time. Beberapa studi relevan telah berkontribusi, seperti penerapan Convolutional Neural Network (CNN) untuk pengenalan alfabet statis [5], kombinasi MediaPipe dengan model sekuensial seperti Long Short-Term Memory (LSTM) untuk tanda dinamis [6], serta pemanfaatan arsitektur deteksi objek seperti YOLOv8 untuk pengenalan langsung melalui kamera [7].

Meskipun penelitian-penelitian tersebut menunjukkan hasil yang menjanjikan, sebagian besar masih menghadapi tantangan fundamental dalam merepresentasikan informasi gestur secara utuh. Model yang ada umumnya mengandalkan data landmark tangan dari MediaPipe dalam bentuk numerik yang diolah secara sekuensial. Pendekatan ini memiliki keterbatasan dalam merepresentasikan hubungan spasial antar titik tangan, yang sebenarnya sangat penting untuk membedakan isyarat dengan bentuk yang serupa.

Hal ini juga ditekankan oleh Gurusiddappa Hugar et al. [8], yang menunjukkan bahwa model sekuensial tradisional

(CNN-LSTM) belum mampu menangkap hubungan spasial dan temporal secara bersamaan dalam pengenalan gesture dinamis. Selain itu, Bader Alsharif et al. [9] melaporkan bahwa sistem interpretasi bahasa isyarat berbasis MediaPipe dan YOLOv11 masih menghadapi kendala dalam kondisi nyata, seperti pencahayaan dan latar belakang yang kompleks, karena sistem ini sangat bergantung pada representasi landmark numerik. Oleh karena itu, dibutuhkan pendekatan baru yang mampu mempertahankan informasi spasial dari titik-titik landmark secara lebih utuh dan informatif, sekaligus meminimalkan pengaruh variabel eksternal [10].

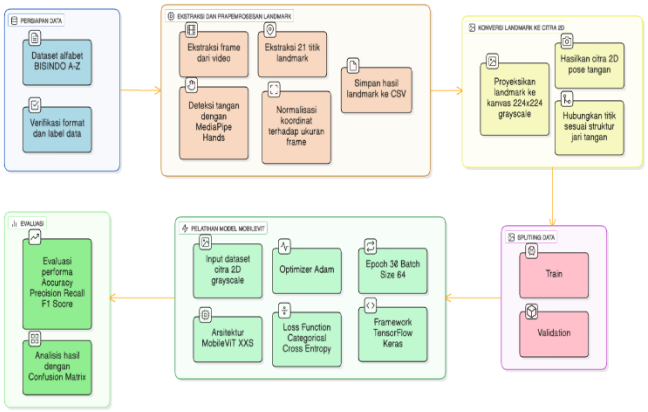
Berbagai penelitian sebelumnya telah menunjukkan hasil yang signifikan dalam pengenalan bahasa isyarat menggunakan MediaPipe Hands, namun sebagian besar masih terbatas pada pengolahan data landmark dalam bentuk koordinat numerik yang diproses secara sekuensial menggunakan model seperti LSTM [11]. Pendekatan tersebut efektif dalam mengenali pola temporal (time-series), tetapi kurang optimal dalam mengekstraksi dan mempelajari hubungan spasial intrinsik antar keypoint tangan. Di sisi lain, beberapa studi yang memanfaatkan citra RGB mentah menghadapi tantangan efisiensi komputasi serta sensitivitas terhadap variasi pencahayaan dan latar belakang yang tinggi [12],[13].

Menjawab keterbatasan tersebut, penelitian ini mengusulkan pendekatan Landmark-to-Image Conversion (L2IC), yaitu metode transformasi koordinat numerik hasil ekstraksi MediaPipe Hands menjadi Peta Titik (Keypoint Map) dua dimensi (2D). Pendekatan L2IC ini secara efektif mengubah data relasional numerik menjadi representasi visual spasial, yang secara langsung mengisolasi pola geometris murni antar titik tangan dari gangguan visual eksternal, masalah utama pada metode berbasis citra piksel penuh. Citra 2D hasil konversi kemudian digunakan sebagai input bagi arsitektur MobileVit-XXS. Model ringan ini dipilih karena menggabungkan keunggulan CNN dalam ekstraksi fitur lokal dan Transformer dalam menangkap hubungan global [14], [15]. Pemilihan MobileViT juga didasari oleh kemampuannya dalam pemrosesan real-time yang efisien dan kompatibilitasnya dengan perangkat mobile dan web [16],[17].

Dengan demikian, penelitian ini bertujuan untuk Mengembangkan metode Landmark-to-Image Conversion (L2IC), Menerapkan arsitektur MobileViT untuk klasifikasi alfabet BISINDO statis, dan Mengevaluasi performa sistem berdasarkan metrik akurasi dan F1-score. Pendekatan yang diusulkan diharapkan dapat menghadirkan sistem penerjemahan BISINDO yang lebih adaptif, ringan, dan akurat, sekaligus memperluas akses komunikasi bagi penyandang tunarungu melalui penerapan teknologi berbasis kecerdasan buatan.

II. METODE

Tahapan kegiatan penelitian ini mengikuti alur kerja yang tersusun secara sistematis, sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Diagram Alur Penelitian

A. Persiapan Data

Tahap ini berfokus pada pengumpulan dan validasi data yang digunakan dalam pelatihan serta pengujian model pengenalan alfabet BISINDO. Dataset yang digunakan dalam penelitian ini merupakan dataset alfabet statis Bahasa Isyarat Indonesia (BISINDO) yang mencakup representasi gestur tangan dari huruf A hingga Z. Data diperoleh dari sumber terbuka yang tersedia pada platform Kaggle [18], disajikan dalam format video yang terstruktur berdasarkan kelas huruf. Setiap folder kelas merepresentasikan satu huruf alfabet dan berisi satu berkas video demonstrasi isyarat tangan yang sesuai.

Setiap video memiliki durasi rata-rata sekitar empat sampai 10 detik dengan format. Sebelum digunakan pada tahap ekstraksi fitur, seluruh video melalui proses verifikasi dan validasi manual untuk memastikan kesesuaian antara label kelas dan gestur tangan yang ditampilkan, serta menjamin konsistensi teknis antar berkas. Tahapan ini penting agar data yang digunakan bebas dari anomali yang dapat memengaruhi performa model pada proses pelatihan dan evaluasi.

Meskipun video berisi sedikit pergerakan transisi menuju pose huruf, frame yang digunakan dalam penelitian ini hanya diambil saat pose tangan telah stabil. Dengan demikian, dataset yang digunakan sepenuhnya merupakan isyarat statis dan tidak mencakup gestur dinamis BISINDO. Untuk memberikan gambaran yang lebih jelas mengenai karakteristik dataset, Tabel 1 merangkum informasi utama terkait jumlah kelas, format data, kondisi perekaman.

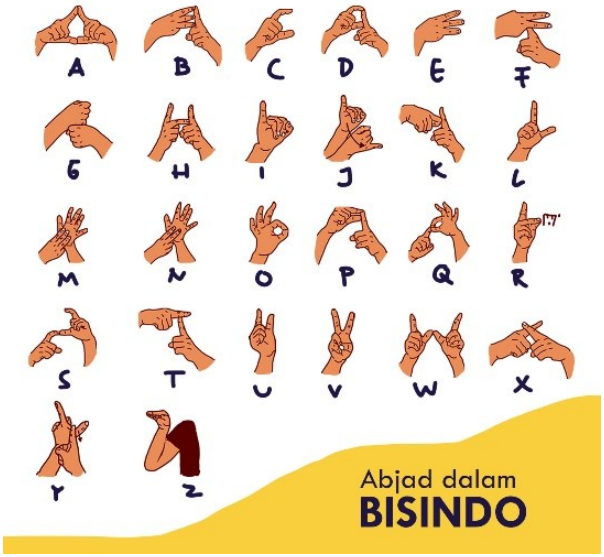
TABEL I
DATASET INFORMATION

Feature	Description
Data Source	Public Dataset (Kaggle: BISINDO Video Dataset)
Number of Classes	26 (A-Z)
Input Data Type (Original)	Video Mp4, Frames (RGB)
Input Data Type (Processed)	JPG
Frame Dimensions (L2IC)	224 x 224 pixels
Total Number of Frames	12.480
Frames per Class	480 frames per class

Number of Signers	2
Recording Conditions	Plain background, stable lighting
Movement Type	Static alphabet
Class Imbalance	No

Meskipun dataset memiliki struktur yang rapi dan konsisten, data yang digunakan masih terbatas dari sisi keragaman pengguna. Seluruh video dalam dataset berasal dari dua signer dengan karakteristik tangan yang relatif homogen, baik dari segi usia, gender, warna kulit, maupun proporsi panjang–pendek jari. Variasi morfologi tangan yang lebih luas, seperti tangan anak-anak, pengguna dengan warna kulit gelap, atau bentuk jari yang berbeda, belum tercakup dalam data pelatihan. Keterbatasan keragaman ini berpotensi memengaruhi generalisasi model ketika diterapkan pada populasi pengguna BISINDO yang lebih beragam.

Sebagai ilustrasi, Gambar 2 menampilkan representasi visual alfabet BISINDO dari huruf A hingga Z, yang menjadi dasar pembentukan kelas-kelas dalam sistem klasifikasi isyarat tangan yang dikembangkan pada penelitian ini.



Gambar 2. Representasi Alfabet BISINDO A-Z

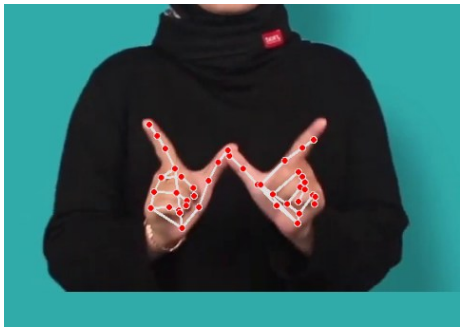
B. *Praprosesan dan Ekstrasi Landmark*

Tahap praproses berperan dalam menyiapkan data agar sesuai dengan format input yang dibutuhkan oleh model MobileVit. Proses ini dilakukan untuk memastikan data bersih dan seragam sebelum memasuki tahap pelatihan. Proses ini meliputi:

1) *Ekstraksi Frame Video*: Seluruh data alfabet Bahasa Isyarat Indonesia (BISINDO) yang digunakan pada penelitian ini berbentuk video berdurasi ±4 detik untuk setiap huruf A–Z. Setiap video diubah menjadi kumpulan citra statis (frame) menggunakan pustaka OpenCV [19]. Tahap ini bertujuan untuk memperoleh representasi visual diskret dari gerakan tangan, meminimalkan efek motion blur, serta menjaga konsistensi antar frame.Setiap frame hasil ekstraksi kemudian digunakan sebagai input pada tahap deteksi tangan. Proses ekstraksi dilakukan dengan membuka video, membaca frame satu per satu, dan menyimpannya ke direktori kerja. Setiap frame diberi label sesuai dengan kelas huruf isyarat yang diwakilinya.

Pada penelitian ini, frame yang digunakan tidak diambil secara acak dari seluruh rangkaian gerakan pada video, melainkan dipilih hanya ketika pose tangan telah stabil dan membentuk huruf alfabet secara penuh. Frame transisi atau pergerakan menuju pose akhir tidak disertakan untuk memastikan bahwa seluruh sampel yang diproses benar-benar mewakili pose statis. Dengan demikian, pipeline pengolahan data ini memanfaatkan satu frame statis untuk setiap sampel tanpa melibatkan rangkaian informasi temporal atau gestur dinamis.

2) *Deteksi dan Pelacakan Tangan:* Setelah kumpulan frame diperoleh, setiap citra diproses menggunakan pustaka MediaPipe Hands untuk mendeteksi dan melacak kedua tangan secara real-time [20]. Algoritma ini mampu mengenali hingga dua tangan secara simultan dengan kestabilan tinggi terhadap variasi pencahayaan dan posisi kamera. MediaPipe secara otomatis mengekstraksi total 42 titik landmark, terdiri atas 21 titik pada tangan kanan dan 21 titik pada tangan kiri, masing-masing dengan tiga koordinat spasial (x, y, z). Proses pelacakan ini divisualisasikan pada Gambar 3,yang menunjukkan hasil deteksi landmark pada kedua tangan dari salah satu frame video.



Gambar 3. Hasil deteksi dan pelacakan dua tangan menggunakan MediaPipe Hands

Meskipun MediaPipe Hands memberikan hasil yang baik pada kondisi ideal, algoritma ini rentan mengalami penurunan performa pada situasi dunia nyata, seperti pencahayaan tidak stabil, occlusion sebagian jari/telapak tangan, serta gerakan tangan yang cepat. Kegagalan ekstraksi landmark pada kondisi tersebut menyebabkan data input ke tahap L2IC menjadi tidak valid dan mengganggu pipeline proses klasifikasi.

3) *Normalisasi dan Penyimpanan Data:* Koordinat landmark hasil deteksi dinormalisasi berdasarkan dimensi frame agar model menjadi invarian terhadap skala dan posisi. Normalisasi ini dilakukan menggunakan Min-Max Scaling per frame pada setiap koordinat. Proses ini memastikan distribusi data yang seragam dan meminimalkan efek

pergeseran tangan dalam frame, di mana koordinat baru dengan rumus:

$$x\_(norm = (x - x\_min)/x\_\max \; 〚 - x\_min 〛)$$

Seluruh hasil ekstraksi kemudian disimpan dalam format Comma-Separated Values (CSV) menggunakan pustaka Pandas. Setiap baris mewakili satu frame, sedangkan setiap kolom menyimpan koordinat landmark tangan kanan dan kiri. Total terdapat 126 fitur numerik (21 × 3 × 2) untuk setiap baris data.Contoh cuplikan struktur data hasil ekstraksi ditunjukkan pada Tabel 2.

TABEL II
DATA HASIL EKSTRAKSI MEDIAPIPE

Class	Right_x1	Right_y1	Right_z1	…	Left_x21	Left_y21	Left_z21
A	0.524	0.388	-0.11	..	0.621	0.374	-0.09
A	0.621	0.449	-0.02	…	0.625	0.576	-0.02

C. Konversi Landmark Ke Citra 2D (L2IC)

Tahap ini berfokus pada transformasi koordinat numerik hasil ekstraksi MediaPipe Hands menjadi citra dua dimensi (2D) berukuran 224 × 224 piksel dalam format grayscale. Representasi ini disebut sebagai Landmark-to-Image Conversion (L2IC), yang bertujuan mempertahankan informasi spasial antar titik tangan dalam bentuk visual sehingga dapat dipelajari secara efektif oleh model MobileViT-XXS. Proses konversi dilakukan melalui tiga tahap utama sebagai berikut.

1) Normalisasi Skala (Min-Max Scaling): Seluruh koordinat x dan y dihitung nilai minimum dan maksimumnya untuk menentukan rentang gerakan tangan:

$$x\_\min \; 〚 = \min(x), 〛 \; x\_\max \; 〚 = \max(x)〛,$$
$$y\_\min \; 〚 = \min(y) 〛 , \; y\_\max \; 〚 = \max(y)〛$$

Nilai rentang ini digunakan untuk menghitung factor skala pada sumbu x dan y agar landmark dapat dipetakan secara proporsional ke dalam kanvas berukuran 224 x 224 piksel dengan margin sebesar 10 piksel:

$$〚scale〛\_(x = (224 - 10)/x\_\max \; 〚 x\_min 〛),$$
$$〚scale〛\_(y = (224 - 10)/y\_\max \; 〚 y\_min 〛)$$

Koordinat yang telah dinormalisasi dihitung dengan:

$$x\_(scaled = (x - x\_\min) \text{ X } 〚 scale 〛\_x),$$
$$y\_(scaled = (y - y\_min) \; X \; 〚 scale 〛\_y)$$

Tahap ini memastikan representasi tangan invariant terhadap ukuran tangan.

2) Translasi Posisi (Offset Centering): Setelah penskalaan, sistem menerapkan offset untuk menempatkan seluruh landmark secara simetris di tengah kanvas:

$$〚offset〛\_(x = (224 - (x\_\max \; 〚 - x\_min 〛) \; X \; 〚 scale 〛\_x)/2)$$
$$〚offset〛\_(y = (224 - (y\_\max \; 〚 - y\_min 〛) X \; 〚 scale 〛\_y)/2)$$

Koordinat akhir dihitung sebagai:

$$x\_(final = x\_scaled + \; 〚 offset 〛\_x),$$
$$y\_(final = y\_(scaled) + \; 〚 offset 〛\_y)$$

Tahap ini membuat gestur invariant terhadap pergeseran posisi tangan pada frame asli.

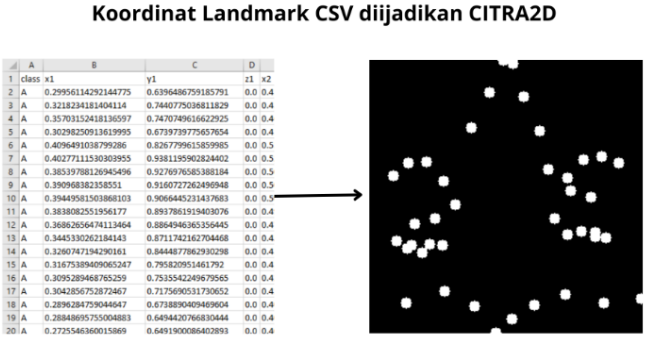
3) Proyeksi ke Citra 2D (Mapping to Image Representation): Koordinat yang telah dinormalisasi kemudia diproyeksikan ke dalam kanvas berukuran 224 x 224 piksel. Setiap titik digambarkan sebagai lingkaran padat berwarna putih(nilai 255) dengan radius 6 piksel:

$$canvas[y\_final, x\_final] = 255$$

Semua titik dipetakan tanpa dihubungkan oleh garis, sehingga menghasilkan representasi point-cloud map yang mengekspresikan pola spasial murni dari bentuk tangan. Pendekatan ini menghilangkan pengaruh latar belakang, pencahayaan, dan warna kulit, serta lebih ringan secara komputasi dibandingkan penggunaan citra RGB.

Pemilihan representasi titik tanpa garis dilakukan secara sengaja untuk menghindari bias struktural yang dapat muncul apabila hubungan antar titik dibentuk secara manual. Menghubungkan titik dengan garis tertentu (misal mengikuti anatomi jari) akan memaksakan asumsi topologi tertentu, padahal MediaPipe sendiri tidak selalu stabil pada urutan landmark ketika terjadi oklusi ringan. Oleh karena itu, pendekatan point-cloud sengaja digunakan agar representasi tetap netral dan hanya memetakan distribusi geometris murni antar titik.

Metode L2IC memberikan keuntungan signifikan dibandingkan representasi numerik sekuensial karena pola spasial antar titik dapat dipelajari secara langsung oleh model vision transformer. Seluruh citra hasil konversi disimpan dalam format PNG dan digunakan sebagai dataset baru pada tahap pelatihan. Proses konversi ditunjukkan pada Gambar 4.



Gambar 4. Konversi Landmark Menjadi Citra 2D Grayscale

D. Splitting Data

Tahap ini merupakan proses pembagian dataset citra hasil konversi L2IC ke dalam dua bagian, yaitu train set dan validation set. Pembagian dilakukan menggunakan pendekatan random per-class split, di mana seluruh citra dalam setiap kelas (A–Z) terlebih dahulu diacak, kemudian dibagi dengan rasio 80% untuk pelatihan dan 20% untuk validasi. Metode ini memastikan bahwa distribusi jumlah sampel per kelas tetap seimbang pada kedua subset.

Namun, pembagian ini belum menerapkan cross-validation, subject-independent split, maupun signer-independent evaluation. Seluruh frame berasal dari dua signer yang sama, dan proses pembagian acak memungkinkan frame dari signer yang sama muncul pada train set dan validation set. Akibatnya, evaluasi terutama mencerminkan intra-signer performance, yaitu kemampuan model mengenali pola tangan dari signer yang juga terlihat pada tahap pelatihan sehingga generalisasi terhadap signer baru belum dapat dinilai secara menyeluruh. Detail pembagian dataset ditampilkan pada Tabel 3.

TABEL III
DETAIL PEMBAGIAN DATASET

Train	***Valid***
12.480	3120

E. Pelatihan ModeL (MobileVit-XXS)

Tahap pelatihan model bertujuan membangun sistem klasifikasi huruf alfabet Bahasa Isyarat Indonesia (BISINDO) dari citra dua dimensi hasil konversi landmark tangan. Arsitektur yang digunakan adalah Mobile Vision Transformer varian ekstra kecil (MobileViT-XXS), yang memadukan keunggulan Convolutional Neural Network (CNN) dalam ekstraksi fitur lokal dan Transformer dalam menangkap hubungan spasial global. Model ini dipilih karena memiliki keseimbangan antara akurasi dan efisiensi komputasi.

Proses pelatihan dilakukan menggunakan dataset citra 2D berukuran 224×224 piksel, dengan parameter utama yang disajikan pada Tabel 4. Model dilatih selama 30 epoch menggunakan batch size 64, learning rate 0,001, dan optimizer Adam dengan label smoothing sebesar 0,1. Fungsi aktivasi Softmax digunakan untuk klasifikasi multi-kelas (26 huruf A–Z).

TABEL IV
HYPERPARAMETER

Parameter	**Value**
Learning rate	0.001
Batch size	64
Epoch	30
Input size	224 x 224 piksel
Patch size	4 x 4
Expansion factor	2
Jumlah kelas	26
Label smoothing	0.1

Model dilatih menggunakan TensorFlow dan Keras pada lingkungan Google Colab Pro dengan bantuan GPU A100.

F. Evaluasi

Tahap evaluasi bertujuan untuk menilai performa model MobileVit-XXS dalam mengenali huruf-huruf alfabet Bahasa Isyarat Indonesia (BISINDO) berdasarkan citra hasil konversi landmark dua dimensi. Evaluasi dilakukan menggunakan data uji yang terpisah dari proses pelatihan untuk memastikan objektivitas hasil dan mengukur kemampuan generalisasi model. Proses evaluasi ini mencakup penghitungan metrik performa utama seperti akurasi (accuracy), presisi (precision), recall, dan F1-score.

Penggunaan metrik F1-score menjadi penting karena mampu menyeimbangkan antara presisi dan recall, terutama ketika terdapat ketidakseimbangan jumlah sampel antar kelas huruf. Metrik-metrik tersebut dihitung menggunakan persamaan sebagai berikut.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$
$$Precision = TP/(TP + FP),\ Recall = TP/(TP + FN)$$
$$F1 - Score = 2\,x\ (Precision\ x\ Recall)/(Precision + Recall)$$

Keterangan: TP (True Positive) adalah jumlah prediksi benar pada kelas positif, TN (True Negative) jumlah prediksi benar pada kelas negatif, FP (False Positive) jumlah prediksi salah pada kelas positif, dan FN (False Negative) jumlah prediksi salah pada kelas negatif.

Selain evaluasi kuantitatif, dilakukan juga analisis Confusion Matrix untuk menggambarkan distribusi prediksi model pada masing-masing kelas huruf alfabet BISINDO. Confusion matrix ini berfungsi untuk mengidentifikasi huruf-huruf yang sering mengalami kesalahan klasifikasi, sehingga dapat dijadikan dasar untuk peningkatan performa model di penelitian selanjutnya.

Sebagai perbandingan, evaluasi juga dilakukan pada tiga konfigurasi model, yakni:

- MLP berbasis data landmark CSV (tanpa L2IC),
- MobileViT-XXS dengan input citra RGB, dan
- MobileViT-XXS dengan input citra hasil konversi L2IC.

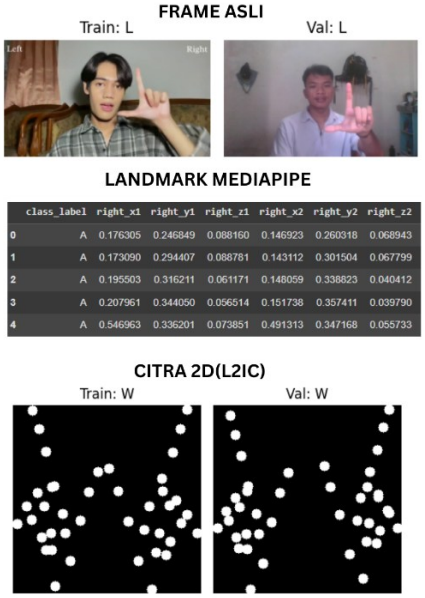
Perbandingan ini memungkinkan analisis yang komprehensif terhadap pengaruh representasi data terhadap performa model, baik dari aspek akurasi maupun efisiensi waktu inferensi.

III. HASIL DAN PEMBAHASAN

A. Analisis Representasi Data L2IC

Seluruh Representasi data merupakan komponen fundamental dalam menentukan efektivitas model pembelajaran mendalam. Penelitian ini mengusulkan metode Landmark-to-Image Conversion (L2IC) untuk mengubah 42

koordinat landmark tangan (21 titik per tangan) dari format numerik (CSV) menjadi Peta Titik dua dimensi dalam format grayscale berukuran 224×224 piksel. Tahapan ini berfungsi sebagai jembatan antara data geometrik spasial hasil MediaPipe dengan model pembelajaran visual MobileViT-XXS.



	class_label	right_x1	right_y1	right_z1	right_x2	right_y2	right_z2
0	A	0.176305	0.246849	0.088160	0.146923	0.260318	0.068943
1	A	0.173090	0.294407	0.088781	0.143112	0.301504	0.067799
2	A	0.195503	0.316211	0.061171	0.148059	0.338823	0.040412
3	A	0.207961	0.344050	0.056514	0.151738	0.357411	0.039790
4	A	0.546963	0.336201	0.073851	0.491313	0.347168	0.055733

Gambar 5. Hasil konversi L2IC dari frame asli menjadi citra Peta Titik 2D.

Visualisasi hasil konversi sebagaimana ditunjukkan Gambar 5, yang memperlihatkan perbandingan antara frame asli, hasil ekstraksi landmark mentah, dan citra Peta Titik hasil konversi L2IC. Proses ini berhasil mengisolasi informasi gestur tangan dari dua individu berbeda, sekaligus menghilangkan gangguan visual seperti variasi pencahayaan, warna kulit, maupun latar belakang. Representasi yang dihasilkan menampilkan himpunan titik putih diskret pada kanvas hitam, yang menggambarkan pola distribusi spasial murni dari posisi tangan yang telah dinormalisasi.

Meskipun proses normalisasi skala dan translasi berhasil menghasilkan representasi yang konsisten, metode L2IC yang digunakan dalam penelitian ini belum menangani variasi rotasi tangan secara eksplisit. Kondisi rotasi ekstrem dapat menyebabkan perubahan distribusi titik yang signifikan, sehingga berpotensi memengaruhi stabilitas pola geometris yang dihasilkan. Selain itu, normalisasi rotasi memerlukan estimasi orientasi telapak tangan (handedness) yang tidak selalu stabil pada MediaPipe ketika terjadi oklusi minor. Oleh sebab itu, rotasi dibiarkan apa adanya dan diperlakukan sebagai limitasi yang perlu ditangani pada penelitian lanjutan

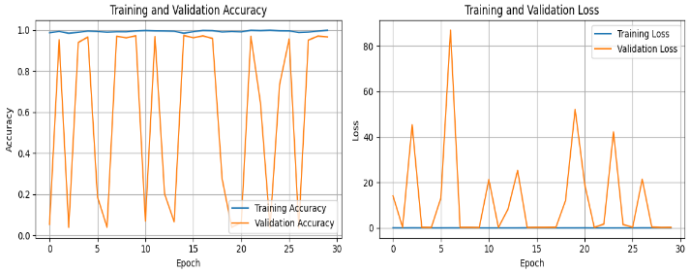
B. Kinerja Klasifikasi MobileVit-XXS Menggunakan L2IC

Tahap ini menyajikan hasil evaluasi kuantitatif performa model MobileViT-XXS dalam tugas pengenalan 26 kelas alfabet Bahasa Isyarat Indonesia (BISINDO). Model diuji menggunakan test set terpisah guna memastikan kemampuan generalisasi yang objektif dan mengukur performa dalam skenario nyata. Evaluasi dilakukan dengan menghitung metrik utama seperti akurasi (Accuracy), presisi (Precision), recall, dan F1-score. Hasil pengujian diringkas dalam Tabel 3.

TABEL V
HASIL EVALUASI MOBILEVIT-XXS DENGAN INPUT L2IC

Parameter Metrik	***Nilai(%)***
Akurasi	97,98
Presisi	98,01
Recall	97,98
F1-Score	97,98

Hasil pengujian pada Tabel 5 menunjukkan bahwa model MobileViT-XXS dengan input citra L2IC (Landmark-to-Image Conversion) menghasilkan performa klasifikasi yang sangat kompetitif. Nilai akurasi rata-rata sebesar 97,98% disertai dengan F1-score identik (97,98%) mengindikasikan keseimbangan yang sangat baik antara presisi dan kemampuan deteksi positif pada seluruh kelas alfabet. Konsistensi antar metrik ini menunjukkan bahwa representasi L2IC yang dinormalisasi secara efektif menjaga invarian spasial antar titik tangan sekaligus mengurangi pengaruh noise visual seperti pencahayaan dan latar belakang..Performa model selama proses pelatihan divisualisasikan melalui grafik Akurasi dan Loss pada Gambar 6.



Gambar 6. Grafik Akurasi dan Loss Pelatihan MobileViT-XXS.

Grafik menunjukkan bahwa training accuracy meningkat secara progresif hingga mendekati nilai maksimum (≈1.0), sementara validation accuracy mengalami fluktuasi pada beberapa epoch awal. Pola ini berkorelasi dengan variasi pada validation loss, yang menandakan adanya perbedaan distribusi antar subset data validasi. Namun, pada epoch ke-30, akurasi validasi stabil di tingkat tinggi, menandakan bahwa teknik regularisasi seperti label smoothing berhasil mengurangi risiko overfitting dan menjaga kemampuan generalisasi model terhadap data baru.

Perlu diakui bahwa Macro F1-Score yang tinggi ini didasarkan pada data keypoint yang diekstrak dari video yang direkam dalam kondisi terkontrol, yang meminimalkan tingkat kegagalan ekstraksi MediaPipe di awal pipeline. Akurasi yang dilaporkan mencerminkan kinerja model dalam kondisi input yang optimal (tidak ada oklusi, pencahayaan stabil).

C. Analisis Komparasi dan Efisiensi

Bagian ini menyajikan hasil evaluasi komparatif antara model yang diusulkan, yaitu MobileViT-XXS dengan representasi L2IC (Landmark-to-Image Conversion) Peta Titik, serta dua konfigurasi pembanding. Pembanding pertama menggunakan Multi-Layer Perceptron (MLP) berbasis data numerik pada format CSV, sedangkan pembanding kedua menggunakan MobileViT-XXS dengan input citra RGB mentah. Evaluasi dilakukan dengan mempertimbangkan macro F1-score dan waktu inferensi yang diukur pada perangkat pengujian. Hasil perbandingan dirangkum pada Tabel 6.

TABEL VI
PERBANDINGAN ANTAR METODE

Model	Macro F1-Score	Waktu Inferensi (s/frame)	Jumlah Parameter (Juta)	Ukuran Model
MLP (CSV)	0,8852	0.001	0.236	920.6 KB
Mobilevit-XXS (L2IC)	0,9798	0.0086	1.367	5,22 MB
MobileVit-XXS (RGB)	0,8682	0. 0029	1.367	5,22 MB

Pengujian menunjukkan bahwa MobileViT-XXS dengan input L2IC Peta Titik memiliki performa terbaik dengan macro F1-score sebesar 0,9798, unggul signifikan dibandingkan MLP dengan skor 0,8852 dan MobileViT-XXS dengan input RGB sebesar 0,8682. Peningkatan performa sebesar ±11,16% dibandingkan baseline RGB mengindikasikan bahwa representasi L2IC mampu mengekstraksi informasi spasial yang lebih murni dengan menekan gangguan visual seperti pencahayaan, warna kulit, serta latar belakang. Hal ini mengkonfirmasi keefektifan representasi berbasis peta titik dalam menggambarkan pola geometris tangan yang krusial untuk pengenalan alfabet BISINDO.

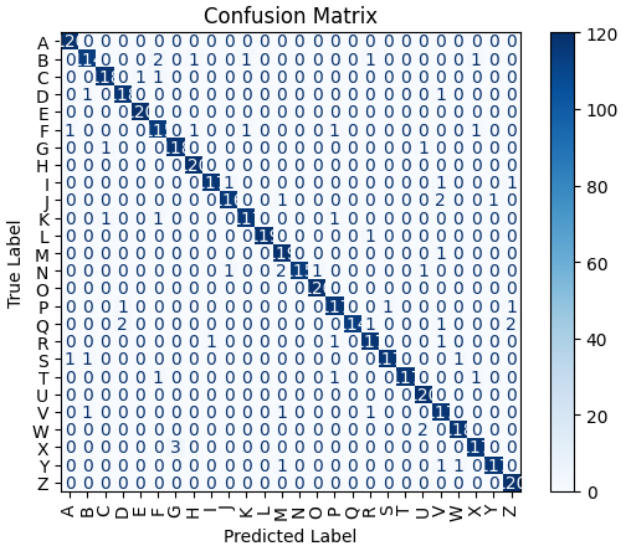
Dari sisi waktu inferensi, MLP mencatat latensi terendah sebesar 0.001 detik per frame, namun akurasinya lebih rendah dibandingkan kedua model MobileViT-XXS. Sementara itu, MobileViT-XXS berbasis L2IC mencatat waktu 0.0029 detik, lebih cepat dibandingkan varian berbasis RGB. Perbedaan ini terutama berasal dari sifat citra grayscale L2IC yang memiliki distribusi piksel lebih sederhana dibandingkan citra RGB.

Namun, penting untuk dicatat bahwa seluruh pengukuran waktu inferensi dilakukan pada perangkat laptop dengan prosesor AMD Ryzen 3 3200U, 20 GB RAM, tanpa GPU eksternal. Dengan demikian, nilai latensi yang dilaporkan hanya menggambarkan performa pada perangkat tersebut dan tidak dapat langsung digeneralisasikan ke perangkat edge seperti Raspberry Pi, Jetson Nano, maupun smartphone kelas menengah ke bawah. Pengujian tambahan pada perangkat dengan kemampuan komputasi lebih rendah diperlukan untuk menilai kelayakan implementasi model pada skenario edge computing.

Secara keseluruhan, hasil komparasi menunjukkan bahwa representasi L2IC memberikan kontribusi positif terhadap peningkatan akurasi model tanpa mengubah struktur arsitektur dasar MobileViT-XXS. Namun, evaluasi lebih lanjut pada berbagai kondisi perangkat perlu dilakukan untuk memastikan karakteristik performa model pada skenario penggunaan yang lebih luas.

D. Analisis Confusion Matrix

Analisis Confusion Matrix dilakukan untuk mengevaluasi pola prediksi model MobileViT-XXS berbasis representasi L2IC (Peta Titik) pada 26 kelas alfabet Bahasa Isyarat Indonesia (BISINDO). Visualisasi hasil pengujian ditampilkan pada Gambar 7.



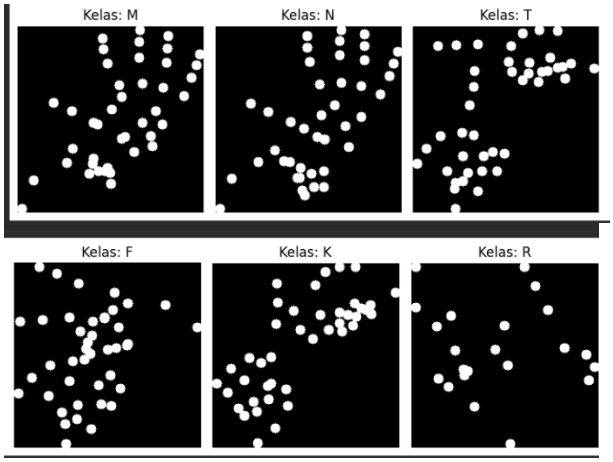
Gambar 7. Confusion Matrix MobileViT-XXS dengan Input L2IC

Sebagaimana terlihat pada Gambar 7, hampir seluruh elemen diagonal utama menunjukkan intensitas warna biru tua yang tegas, menandakan bahwa model memiliki tingkat prediksi benar yang sangat tinggi untuk sebagian besar huruf alfabet. Pola diagonal yang dominan ini mengindikasikan bahwa sistem berhasil mengenali gestur tangan dengan konsistensi yang cukup baik di sebagian kelas.

TABEL VII
DAFTAR KESALAHAN KLASIFIKASI

Pasangan Huruf Yang Tertukar	Kemungkinan Penyebab Kesalahan
M ↔ N	Kemiripan posisi jari tengah dan telunjuk; perbedaan hanya pada jumlah jari yang dilipat sebagian.
T ↔ F	Pose jari telunjuk dan ibu jari yang tumpang tindih secara visual.
K ↔ R	Sudut rotasi pergelangan tangan dan variasi arah pandang kamera.

Meskipun demikian, terdapat beberapa kesalahan klasifikasi minor yang terjadi pada huruf-huruf dengan kemiripan pose tangan tinggi. Contohnya, huruf M dan N sering tertukar akibat kemiripan posisi jari, sementara huruf T dan F menunjukkan kesalahan karena tumpang tindih visual antara jari telunjuk dan ibu jari. Kesalahan lain juga ditemukan pada pasangan K dan R, yang disebabkan oleh variasi kecil pada rotasi pergelangan tangan dan sudut pandang kamera. Pola kesalahan tersebut dirangkum pada Tabel 7.



Gambar 8. Visualisasi L2IC pada kelas yang sering tertukar (M–N, T–F, K–R)

Sebagai pelengkap analisis kesalahan, Gambar 8 menampilkan contoh representasi L2IC dari huruf-huruf yang menjadi sumber kesalahan utama. Terlihat bahwa pola geometris pada pasangan seperti M–N, T–F, dan K–R memiliki distribusi titik yang sangat mirip, sehingga wajar menjadi penyebab utama mis-klasifikasi. Visualisasi ini menguatkan temuan dari Confusion Matrix dan menjelaskan penyebab kesalahan secara lebih jelas.

Meskipun model menunjukkan performa yang sangat tinggi pada pengenalan alfabet BISINDO berbasis representasi L2IC, terdapat beberapa keterbatasan yang perlu diperhatikan. Dataset yang digunakan sepenuhnya terdiri dari pose statis, sehingga sistem belum mampu mengenali gestur dinamis yang merupakan bagian penting dari BISINDO dalam situasi komunikasi nyata. Frame yang diproses juga hanya diambil ketika pose tangan telah stabil, sehingga informasi transisi gerakan tidak dilibatkan dalam proses pelatihan. Selain itu, ketergantungan pada MediaPipe untuk ekstraksi landmark belum diuji pada kondisi dunia nyata seperti pencahayaan rendah, oklusi jari, atau pergerakan tangan cepat.

Dataset yang relatif homogen dari segi latar belakang, variasi pengguna, serta karakteristik tangan juga dapat membatasi kemampuan generalisasi model ketika digunakan oleh populasi pengguna yang lebih beragam.

Selain itu, seluruh pengujian pada penelitian ini dilakukan secara offline dalam kondisi terkontrol sehingga belum mencerminkan skenario penggunaan dunia nyata. Sistem belum diuji pada gestur dengan gerakan cepat, tangan sebagian tertutup, latar belakang kompleks, penggunaan dua tangan secara simultan, ataupun kondisi pencahayaan yang berubah-ubah. Penelitian ini juga belum menghitung end-to-end latency yang meliputi deteksi keypoint, proses L2IC, dan inferensi model. Oleh karena itu, performa waktu nyata belum dapat digeneralisasikan dan hasil akurasi yang tinggi masih bersifat optimistik. Keterbatasan-keterbatasan ini menjadi dasar penting untuk penelitian lanjutan, terutama pengujian dalam skenario dunia nyata dan validasi performa sistem secara end-to-end pada kondisi penggunaan publik.

IV. KESIMPULAN

Penelitian ini berhasil mengembangkan dan mengevaluasi pendekatan hibrida Landmark-to-Image Conversion (L2IC) yang diintegrasikan dengan arsitektur MobileViT-XXS untuk tugas klasifikasi alfabet statis Bahasa Isyarat Indonesia (BISINDO). Hasil eksperimen menunjukkan bahwa representasi L2IC dalam bentuk Peta Titik mampu mempertahankan relasi spasial antar keypoint secara efektif dan menghasilkan kinerja klasifikasi yang tinggi, dengan nilai Macro F1-Score sebesar 0,9798, atau meningkat sekitar 11,16% dibandingkan baseline MobileViT-XXS berbasis citra RGB (0,8682).Waktu inferensi rata-rata sebesar 0,0029 detik per frame menggambarkan performa komputasi pada perangkat uji (laptop dengan CPU AMD Ryzen 3 3200U), namun tidak dapat digeneralisasikan sebagai performa untuk seluruh perangkat edge. Dengan demikian, penelitian ini tidak mengklaim implementasi real-time maupun efisiensi pada edge devices.

Analisis Confusion Matrix menunjukkan bahwa kesalahan utama terjadi pada huruf dengan kemiripan geometris tinggi seperti M–N dan T–F, yang disebabkan oleh distribusi keypoint yang sangat mirip dan variasi fleksi jari. Selain itu, pipeline sangat bergantung pada keberhasilan ekstraksi landmark oleh MediaPipe Hands, yang belum diuji pada kondisi dunia nyata seperti pencahayaan buruk, oklusi jari, latar belakang kompleks, maupun gerakan tangan cepat. Seluruh pengujian juga dilakukan secara offline dan pada pose statis dalam lingkungan terkontrol, sehingga hasil akurasi yang tinggi ini masih bersifat optimistik serta belum mewakili performa sistem ketika digunakan dalam komunikasi BISINDO yang bersifat dinamis dan tidak terstruktur.

Penelitian lanjutan disarankan untuk mengevaluasi *robustness* sistem pada kondisi dunia nyata, mengukur *end-to-end latency* (deteksi keypoint + L2IC + inferensi) melalui benchmarking dengan model lightweight terkemuka (seperti MobileNetV3 dan EfficientNet-Lite), serta melakukan validasi terhadap dataset dengan variasi signer, latar belakang, dan karakteristik tangan yang lebih beragam. Pengembangan berikutnya juga dapat mengintegrasikan informasi koordinat Z dan pemodelan temporal untuk

memungkinkan pengenalan gestur dinamis pada tingkat kata maupun kalimat.

DAFTAR PUSTAKA

[1] "Potret Penyandang Disabilitas di Indonesia: Hasil Long Form SP2020 - Badan Pusat Statistik Indonesia." Accessed: Oct. 15, 2025.[Online].Available: https://www.bps.go.id/id/publication/2024/12/20/43880dc0f8be5ab92199f8b9/potret-penyandang-disabilitas-di-indonesia--hasil-long-form-sp2020.html

[2] "World report on hearing." Accessed: Oct. 15, 2025. [Online]. Available: https://www.who.int/publications/i/item/9789240020481

[3] F. Tsina, A. Kusmawati, J. K. Ahmad Dahlan, K. Ciputat Timur, and K. Tangerang Selatan, "Dukungan Sosial Terhadap Kualitas Hidup Kelompok Tuli Di Majelis Ta'lim Tuli Indonesia," *Huk. Inov. J. Ilmu Huk. Sos. dan Hum.*, vol. 1, no. 2, pp. 71–80, Mar. 2024, doi: 10.62383/HUMIF.V1I2.94.

[4] D. A. Saraswati, V. D. Towidjojo, and Hasanuddin, "Bahasa Isyarat Indonesia," *J. Med. Prof.*, vol. 4, no. 1, pp. 8–14, 2022, Accessed: Oct. 15, 2025. [Online]. Available: https://jurnal.fk.untad.ac.id/index.php/medpro/article/view/582

[5] E. L. Kelana, M. R. A. Prasetya, . M., and M. Zulfadhilah, "Integrating the CNN Model with the Web for Indonesian Sign Language (BISINDO) Recognition," *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 883–896, Jun. 2025, doi: 10.30871/JAIC.V9I3.9345.

[6] M. Z. Uddin, C. Boletsis, and P. Rudshavn, "Real-Time Norwegian Sign Language Recognition Using MediaPipe and LSTM," *Multimodal Technol. Interact. 2025, Vol. 9, Page 23*, vol. 9, no. 3, p. 23, Mar. 2025, doi: 10.3390/MTI9030023.

[7] B. Alsharif, E. Alalwany, and M. Ilyas, "Transfer learning with YOLOV8 for real-time recognition system of American Sign Language Alphabet," *Franklin Open*, vol. 8, p. 100165, Sep. 2024, doi: 10.1016/J.FRAOPE.2024.100165.

[8] G. Hugar, R. M. Kagalkar, and A. Das, "Comparative Study of Hybrid Deep Learning Models for Kannada Sign Language Recognition," *Int. J. Comput. Intell. Syst.*, vol. 18, no. 1, pp. 1–23, Dec. 2025, doi: 10.1007/S44196-025-00922-4/TABLES/7.

[9] B. Alsharif, E. Alalwany, A. Ibrahim, I. Mahgoub, and M. Ilyas, "Real-Time American Sign Language Interpretation Using Deep Learning and Keypoint Tracking," *Sensors 2025, Vol. 25, Page 2138*, vol. 25, no. 7, p. 2138, Mar. 2025, doi: 10.3390/S25072138.

[10] S. Suherman, A. Suhendra, and E. Ernastuti, "Method Development Through Landmark Point Extraction for Gesture Classification With Computer Vision and MediaPipe," *TEM J.*, vol. 12, no. 3, pp. 1677–1686, Aug. 2023, doi: 10.18421/TEM123-49.

[11] S. Kamble, "SLRNet: A Real-Time LSTM-Based Sign Language Recognition System," Jun. 2025, Accessed: Oct. 15, 2025. [Online]. Available: https://arxiv.org/pdf/2506.11154

[12] D. Amalfitano, V. D. ' Angelo, A. M. Rinaldi, C. Russo, and C. Tommasino, "Enhancing Gesture Recognition for Sign Language Interpretation in Challenging Environment Conditions: A Deep Learning Approach.," *pdfs.semanticscholar.orgD Amalfitano, V D'Angelo, AM Rinaldi, C Russo, C TommasinoKDIR, 2023•pdfs.semanticscholar.org*, 2023, doi: 10.5220/0012209700003598.

[13] J. Qin, M. W.-S. Reports, and undefined 2025, "Sign language recognition based on dual-channel star-attention convolutional neural network," *nature.comJ Qin, M WangScientific Reports, 2025•nature.com*, Accessed: Oct. 15, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-13625-9

[14] S. Mehta and M. Rastegari, "Mobilevit: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer," *ICLR 2022 - 10th Int. Conf. Learn. Represent.*, vol. 3, 2022.

[15] K. Meng and K. Chen, "Navigating Efficiency in MobileViT through Gaussian Process on Global Architecture Factors," Jun. 2024, Accessed: Oct. 15, 2025. [Online]. Available: https://arxiv.org/pdf/2406.04820

[16] F. De *et al.*, "Reducing Computational Cost in MobileViT for Edge-Oriented Models Through Token Merging," *Electron. 2024, Vol. 13, Page 5009*, vol. 13, no. 24, p. 5009, Dec. 2024, doi: 10.3390/ELECTRONICS13245009.

[17] M. Zhang *et al.*, "Dual-Attention-Enhanced MobileViT Network: A Lightweight Model for Rice Disease Identification in Field-Captured Images," *Agric.*, vol. 15, no. 6, pp. 1–22, 2025, doi: 10.3390/agriculture15060571.

[18] "Bisindo - Video Dataset." Accessed: Oct. 15, 2025. [Online]. Available: https://www.kaggle.com/datasets/rizkyyangpalsu/bisindo-video-dataset

[19] "Learning OpenCV: Computer Vision with the OpenCV Library - Gary Bradski, Adrian Kaehler - Google Books." Accessed: Oct. 15, 2025.

[20] F. Zhang *et al.*, "MediaPipe Hands: On-device Real-time Hand Tracking," 2020, [Online]. Available: http://arxiv.org/abs/2006.10214