Buletin Ilmiah Sarjana Teknik Elektro
Vol. 8, No. 3, June 2026, pp. 811-823
ISSN: 2685-9572, DOI: 10.12928/biste.v8i3.15587

Real-Time BISINDO Alphabet Recognition via Faster R-CNN Incorporating Skin Tone Diversity as a Classification Feature

Lilis Nur Hayati [1,3], Anik Nur Handayani [1], Wahyu Sakti Gunawan Irianto [1], Rosa Andrie Asmara [2], Dolly Indra [3], Nor Salwa Damanhuri [4]

[1] Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia
[2] Information Technology Department, Politeknik Negeri Malang, Malang, Indonesia
[3] Department of Computer Science, Universitas Muslim Indonesia, Makassar, Indonesia
[4] Electrical Engineering Studies, Universiti Teknologi MARA (UiTM), Cawangan Pulau Pinang, Malaysia

ARTICLE INFORMATION

Article History:

Received 10 December 2025
Revised 28 March 2026
Accepted 15 June 2026

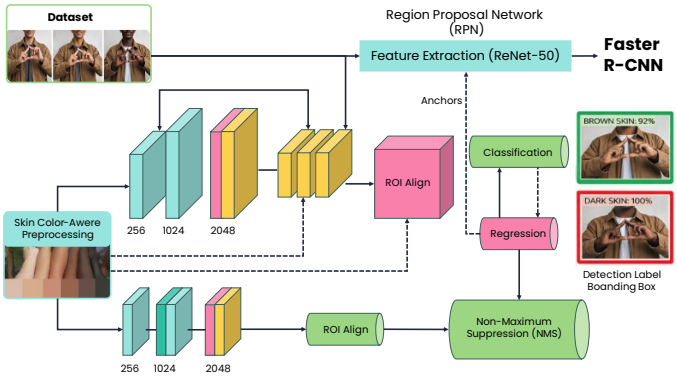
Keywords:

BISINDO;
Faster R-CNN;
Skin Color Features;
Hand Gesture Recognition;
Assistive Technology

Corresponding Author:

Anik Nur Handayani,
Department of Electrical Engineering and Informatics,
Universitas Negeri Malang,
Malang, Indonesia.
Email: aniknur.ft@um.ac.id

ABSTRACT



Indonesian Sign Language (Bahasa Isyarat Indonesia/BISINDO) enables communication for deaf individuals through hand gestures, yet limited public awareness creates significant barriers between deaf and hearing communities. Existing recognition systems often fail to generalize across diverse skin tones, reducing their effectiveness in inclusive real-world deployment. The contribution of this research is a BISINDO alphabet recognition system that integrates skin color features - extracted via HSV-based skin segmentation - as an additional preprocessing layer within the Faster R-CNN framework, explicitly improving detection robustness across varied skin tones. The dataset consists of 8,000 images from ten adult actors representing light, medium-brown, and dark skin tones, augmented through flipping and brightness variation, with a 90:10 training-to-testing ratio. The model was trained over 15,000 steps with a batch size of 24, selected through empirical validation to balance convergence stability and dataset size. Experimental results show that indoor conditions outperform outdoor settings due to controlled lighting. Light-skinned and dark-skinned participants achieved the highest accuracy of 87.5% and F1-score of 85.71%, while medium-brown-skinned participants showed slightly lower performance, likely attributed to greater variability in reflectance under mixed lighting. The system achieves 24 frames per second, demonstrating potential for real-time communication support. These findings confirm that Faster R-CNN with skin color feature integration is effective for BISINDO alphabet recognition, with skin tone diversity being a critical performance factor. Future work will explore larger participant pools and dynamic gesture recognition under varied real-world lighting scenarios.

Document Citation:

L. N. Hayati, A. N. Handayani, W. S. G. Irianto, R. A. Asmara, D. Indra, and N. S. Damanhuri, "Real-Time BISINDO Alphabet Recognition via Faster R-CNN Incorporating Skin Tone Diversity as a Classification Feature," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 8, no. 3, pp. 811-823, 2026, DOI: 10.12928/biste.v8i3.15587.

1. INTRODUCTION

Effective communication is a fundamental aspect of social interaction, enabling individuals to express ideas, emotions, and intentions [1]. For people with hearing impairments, sign language serves as the primary communication medium, utilizing structured hand gestures, body postures, and facial expressions to convey meaning [2][3]. In Indonesia, the deaf community predominantly uses Bahasa Isyarat Indonesia (BISINDO), a naturally developed sign language widely adopted in daily communication [4][5]. Despite its importance, communication barriers between deaf and hearing communities remain significant due to limited public awareness of BISINDO, restricting social participation for deaf individuals [6]. This condition underscores the urgent need for assistive technologies capable of translating BISINDO gestures into textual information to support more inclusive communication [7].

A critical yet frequently overlooked challenge in vision-based sign language recognition is skin color variation [8]. Differences in melanin levels and illumination conditions significantly affect image appearance and feature extraction, causing performance inconsistencies across users with different skin tones [9][10]. Existing recognition systems are often trained on datasets with homogeneous subjects, leading to dataset bias that reduces model generalization, particularly under outdoor or low-light conditions [11][12]. Furthermore, many current approaches focus exclusively on hand shape or motion features without systematically accounting for skin tone diversity, limiting their fairness and reliability in real-world deployment [13]. These gaps directly motivate the present study.

Recent advances in deep learning have enabled substantial progress in gesture recognition and object detection. Convolutional Neural Networks (CNNs) have been widely applied to hand gesture recognition due to their capacity to automatically learn discriminative visual features from image data [14][15]. Among detection frameworks, region-based architectures that integrate region proposal mechanisms with convolutional backbones have demonstrated strong performance across diverse domains including medical imaging, agriculture, and human activity recognition [16][17]. Several studies have explored sign language recognition using these deep learning approaches under controlled environments, reporting promising accuracy [18][19]. However, most existing works do not address skin tone diversity or real-world environmental variability, leaving a significant robustness gap [20][21].

For the BISINDO recognition task, this study selects Faster R-CNN as the detection framework [20], [22][23]. While single-stage detectors such as YOLOv8 and SSD MobileNet offer faster inference, Faster R-CNN's two-stage architecture comprising a Region Proposal Network (RPN) that generates candidate object regions, followed by Region of Interest (RoI) pooling and classification-provides superior localization accuracy for hand gesture regions, which is critical when detecting overlapping or complex hand configurations in BISINDO alphabets [24][25]. Furthermore, Faster R-CNN's modular design allows explicit integration of supplementary features, such as skin color, into the detection pipeline, making it more suitable for this study's objective of analyzing cross-skin-tone performance. The system is designed to achieve real-time processing at 24 frames per second, demonstrating its viability for practical communication support applications.

To address the identified gaps, this study incorporates HSV-based skin color features as an explicit preprocessing component within the Faster R-CNN framework and evaluates the system using a diverse dataset representing light, medium-brown, and dark skin tones under both indoor and outdoor conditions [26]. It is hypothesized that explicitly incorporating skin color features will improve detection consistency across skin tone groups, thereby increasing model robustness and inclusivity. The contribution of this research is the development of a BISINDO alphabet recognition system that integrates skin color features into a Faster R-CNN framework, systematically evaluated across diverse skin tones and environmental conditions to provide empirical insights into the robustness and fairness of deep learning–based sign language recognition systems

2. THEORETICAL FOUNDATION OF FASTER R-CNN

The proposed system architecture is illustrated in Figure 1, which depicts the complete pipeline from raw image input to final alphabet prediction for BISINDO gestures, with skin-tone-aware preprocessing integrated at the front end [18]. Input frames are captured from actors spanning the three skin tone groups during gesture performance. Prior to entering the main detection network, each image undergoes a preprocessing step designed to normalize the effects of lighting variation without modifying the core architecture of Faster R-CNN, given that photometric conditions and pigmentation interact to produce appearance shifts that can compromise detection consistency [27][28]. The normalized frames are subsequently routed through a ResNet-50 convolutional backbone, which produces a hierarchy of feature representations encoding geometric properties such as boundary contours, surface texture patterns, and the spatial arrangement of hand structures [29][30].

Feature maps output by the backbone are simultaneously fed into the RPN, whose anchor-based scanning mechanism proposes candidate bounding boxes—referred to as Regions of Interest (RoIs)—that may enclose gesture-performing hand regions [31][32]. An RoI Align operation then resamples each proposal region onto a fixed spatial grid while preserving subpixel alignment, which is critical for accurate localization [33]. The resampled features advance to a dual-head detection module: one head assigns each proposal to a gesture class (or background), while the other refines the bounding box coordinates through regression. A final Non-Maximum Suppression (NMS) pass eliminates spatially redundant high-confidence predictions, yielding detections labeled with the BISINDO alphabet class, bounding box, and confidence score [34]. The resulting pipeline supports consistent and accurate gesture decoding across the full range of tested skin tones and lighting environments.

The design choice to integrate Faster R-CNN with skin-tone-aware class definitions responds to documented challenges in deploying gesture recognition under real-world conditions, including inconsistent illumination, cluttered backgrounds, and heterogeneous user populations. Research in the field of algorithmic fairness has established that recognition systems trained on phenotypically homogeneous data tend to underperform for under-represented groups, particularly when testing environments introduce photometric variability [23]. Evidence from hardware-constrained deployment studies further underscores the importance of building detection models that balance accuracy with computational efficiency [35]. Region proposal frameworks have been repeatedly validated for their superior spatial localization in complex visual scenes, making them well suited for articulated hand gesture tasks [36].

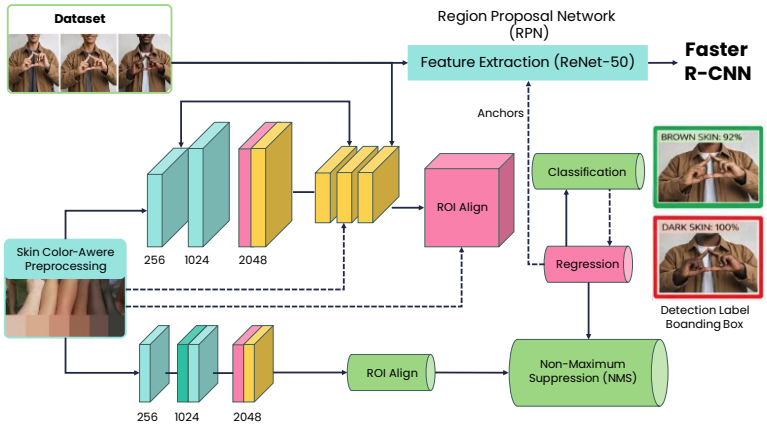


Figure 1. The architecture of Faster R-CNN

3. METHODS

The research design comprises two main phases: the training phase and the testing phase. The training hase covers dataset preparation and model development, while the testing phase evaluates the trained model under real-time conditions. The overall research workflow is illustrated in the flowchart in Figure 2.

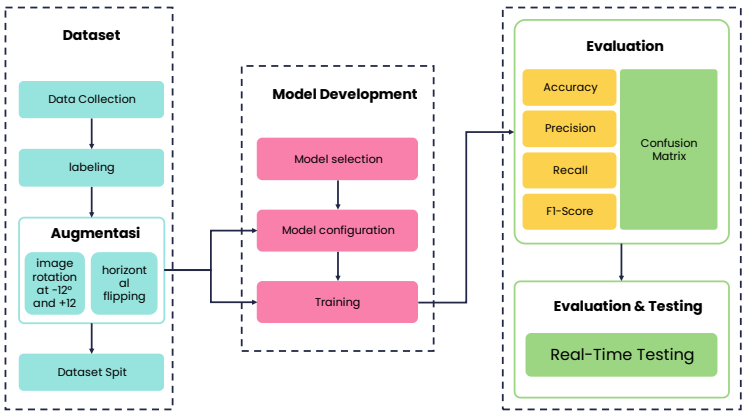


Figure 2. Research Methodology Flowchart

3.1. Dataset Preparation

3.1.1. Dataset Collection

Image data were collected using a smartphone camera from ten adult actors demonstrating eight BISINDO alphabet classes (A–H). The actors represented three skin color categories: four light-skinned, three medium-brown-skinned, and three dark-skinned individuals. Data acquisition was performed at a distance of approximately 70 cm, with the camera positioned at chest level (120–135 cm from the floor) from a frontal view [37]. Captures were conducted under both indoor and outdoor conditions. Indoor illumination ranged from 161 lux to 254 lux, while outdoor illumination ranged from 197 lux to 6,536 lux.

A total of 8,000 original images were collected, yielding approximately 800 images per actor across eight gesture classes. All images were manually annotated using the LabelImg application [38], and each image was resized to 640×640 pixels to match the Faster R-CNN input resolution. Representative samples are illustrated in Figure 3.

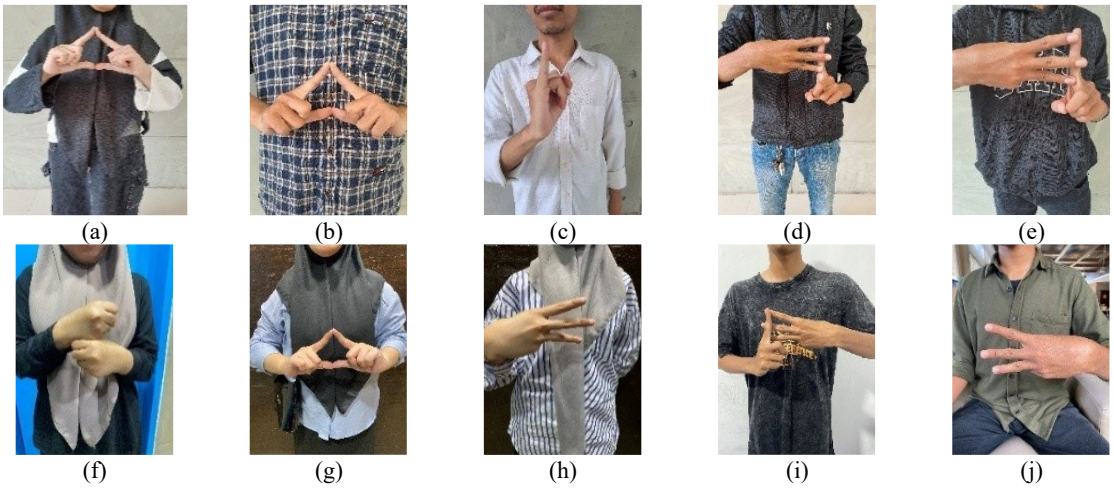


Figure 3. Representative BISINDO Gesture Samples. Examples of BISINDO alphabet gestures (A–H) performed by actors with (a–d) light skin, (e–g) medium-brown skin, and (h–j) dark skin tones under indoor and outdoor conditions with heterogeneous backgrounds

3.1.2. Data Augmentation

To increase dataset diversity and reduce overfitting, data augmentation was applied using the Roboflow platform. Augmentation techniques included horizontal flipping and image rotation at −12° and +12°, chosen to simulate realistic variations in hand orientation. Through this process (Table 1), the original 8,000 images were expanded to 46,916 images. The augmented dataset was exported in TensorFlow Record (TFRecord) format. The dataset was split into 42,259 training images and 4,657 testing images (90:10 ratio). Examples of augmentation results are shown in Figure 4 and Figure 5.

Table 1. Dataset Composition Before and After Augmentation

Condition	Original Images	Augmented Images	Total
Indoor	4,000	19,458	23,458
Outdoor	4,000	19,458	23,458
Total	8,000	38,916	46,916

Original Flip horizontal

Figure 4. Augmentation Result — Horizontal Flip

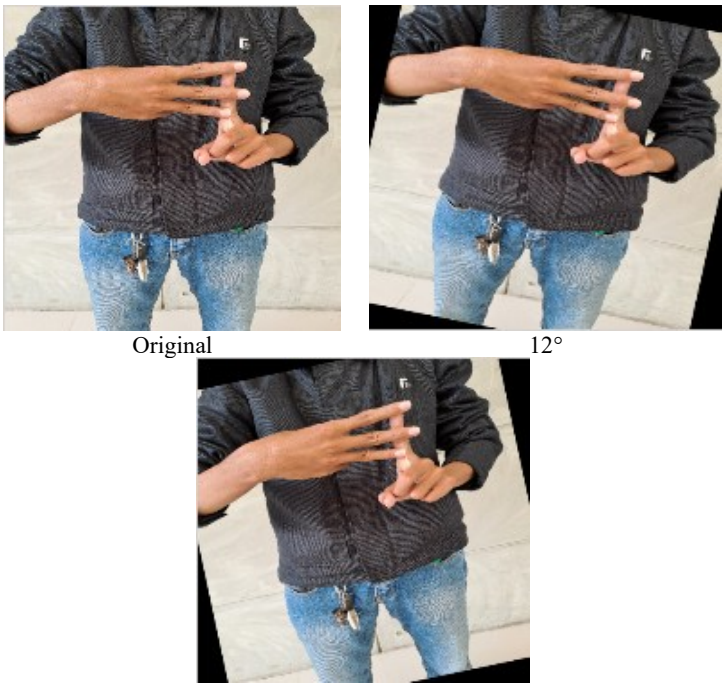


Figure 5. Augmentation Result — Rotation

3.2. Model Development

3.2.1. Skin Color Feature Integration Strategy

A key contribution of this study is the explicit integration of skin color as a discriminative factor within the Faster R-CNN detection framework. This integration was implemented through a **class-definition strategy**: the detection model was configured to recognize 24 classes, representing all combinations of 8 BISINDO alphabet gestures (A–H) and 3 skin color categories (light, medium-brown, dark).

The class-definition strategy differs from explicit skin segmentation approaches (e.g., HSV or YCbCr masking) in two important ways. First, it avoids segmentation failure cases that arise when background colors closely resemble skin tones — such as wooden walls or cream-colored clothing. Second, compared to standard augmentation-only approaches that randomly apply brightness and color jitter without demographic grounding, the class-level skin tone encoding directly supervises the model to learn gesture representations tied to specific skin tone conditions, providing a structured form of domain diversity that augmentation alone cannot guarantee [39][40]. This is evidenced by the consistent 100% recall across all skin tone groups.

3.2.2. Model Selection

As shown in Table 2, several pre-trained models from the TensorFlow 2 Object Detection Model Zoo were evaluated based on inference speed and COCO mAP [41]. The Faster R-CNN ResNet-50 V1 640×640 was selected for its favorable balance between inference speed and detection accuracy. Compared to heavier backbones such as VGG16 (approximately 138 million parameters), ResNet-50 offers a substantially more efficient parameter structure [42]. Although single-stage detectors such as YOLOv8 offer faster raw inference, Faster R-CNN's two-stage architecture — comprising a RPN for candidate region generation followed by RoI pooling and classification — provides superior localization accuracy for complex hand gesture configurations [24]. A direct controlled comparison with YOLOv8 and SSD MobileNet on the identical dataset is identified as future work.

Table 2. Pre-Trained Model Comparison from TensorFlow 2 Model Zoo

No	Model Name	Input Size	Speed (ms)	COCO mAP
1	Faster R-CNN ResNet50 V1 ★	640×640	53	29.3
2	Faster R-CNN ResNet50 V1	1024×1024	65	31.0
3	Faster R-CNN ResNet50 V1	800×1333	65	31.6
4	SSD ResNet101 V1 FPN	1024×1024	104	39.5
5	SSD ResNet152 V1 FPN	1024×1024	111	39.6

3.2.3. Model Training Configuration

Training (Figure 6) was conducted on a system equipped with Intel Core i7 processor, 16 GB RAM, and NVIDIA GPU with CUDA support using Jupyter Notebook version 7.2.1. As shown in Table 3 and Table 4, the training loss decreased consistently from 0.4249 at step 500 to 0.192 at step 15,000, indicating progressive learning of discriminative gesture features.

Table 3. Model Training Hyperparameters

Parameter	Value
Number of detection classes	24 (8 alphabets × 3 skin tones)
Batch size	24
Training steps	15,000
Learning rate (initial)	0.0001
Optimizer	Momentum SGD
Input resolution	640 × 640 pixels
Backbone	ResNet-50 V1

Table 4. Training Loss Progression

Training Step	Training Loss	Elapsed Time
500	0.4249	28 min 9 sec
2,800	0.3389	3 hr 28 min
5,000	0.2868	5 hr 58 min
10,000	0.2267	12 hr 2 min
15,000	0.1920	18 hr 5 min

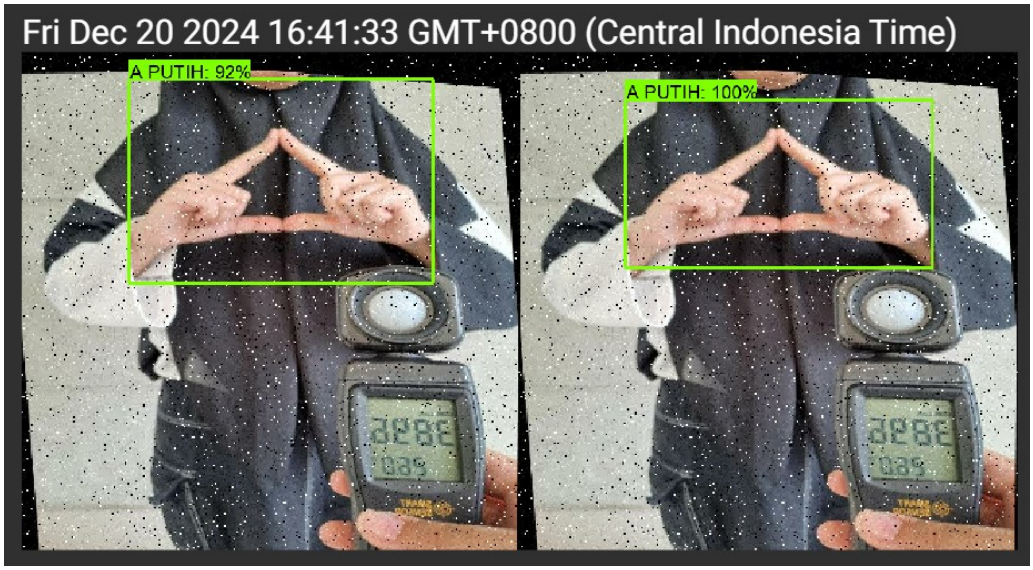


Figure 6. Detection Result Example — BISINDO Alphabet "A"

3.3. Faster R-CNN Detection Architecture

Upon receiving an input frame, the ResNet-50 backbone maps the image to multi-scale feature representations through its convolutional layers. These spatial feature maps are simultaneously shared with the RPN, which uses predefined anchor templates at multiple scales to scan for candidate object locations and outputs a ranked set of region proposals. Each proposal undergoes RoI Align resampling to produce fixed-dimension pooled features, which are then processed by the classification head assigning a gesture class label and the regression head refining the bounding box geometry. Overlapping detections above the confidence threshold are pruned via NMS, and the surviving predictions constitute the model output: an alphabet label, a bounding box, and an associated confidence score [23],[26],[43].

3.4. Testing Pipeline with Testing Protocol

Performance assessment employed two complementary approaches. The first applied the trained model to 4,657 held-out test images, computing confusion-matrix-derived metrics accuracy, precision, recall, and F1-score—for each skin tone and environment combination [44]. The second conducted real-time webcam-based evaluation, with five repetitions per testing condition (self-testing, third-party user testing, varied background

testing) at distances between 30 cm and 70 cm. Detections were accepted only when confidence scores exceeded an empirically determined threshold of 0.5, selected to balance detection sensitivity against false positive rate.

3.5. Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, and F1-score derived from the confusion matrix. Figure 7 illustrates example values for each confusion matrix element.

Figure 7. Illustration of TP, TN, FP, and FN in Gesture Detection Evaluation

4. RESULT AND DISCUSSION

Recognition outcomes are examined across the three skin tone categories and both testing environments, with emphasis on how performance responds to differences in illumination and photometric background complexity.

4.1. Result Overall Performance Across Skin Tone Categories and Environments

Table 5 presents the comprehensive recognition metrics for all combinations of skin tone category and testing environment. A consistent pattern emerges: indoor evaluations outperform their outdoor counterparts across every skin tone group, attributable to the greater photometric stability of interior environments. The highest recognition scores—accuracy of 87.5% and F1-score of 85.71%—were achieved for both light-tone and dark-tone actors tested under indoor conditions.

Across every tested scenario, recall reached 100%, indicating that the model did not fail to detect any target gesture when it appeared in the frame. Precision values were consistently lower than recall, revealing those false positive detections—instances where the model incorrectly identified background regions or non-target gestures as valid alphabet signs—constitute the dominant error mode [45]. Crucially, the inter-group performance spread across skin tone categories was confined to a maximum of 4.5 percentage points, providing quantitative evidence that the skin-tone-encoding strategy did not produce systematic bias favoring any particular demographic subgroup [46].

Table 5. Recognition Performance Summary Across Skin Tone Categories and Testing Conditions

Testing Condition	Skin Tone	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Indoor	Light	87.5	75.0	100.0	85.71
Indoor	Medium-Brown	83.0	66.7	100.0	80.0
Indoor	Dark	87.5	75.0	100.0	85.71
Outdoor	Light	83.0	66.7	100.0	80.0
Outdoor	Medium-Brown	83.0	66.7	100.0	80.0
Outdoor	Dark	83.0	66.7	100.0	80.0

4.2. Confusion Matrix Examples

Figure 8 captures the confusion matrix for the most challenging evaluation scenario: outdoor testing with medium-brown-tone actors, yielding accuracy 83.0%, precision 66.7%, recall 100%, and F1-score 80.0%. Figure 9 presents the peak-performance matrix for indoor light-tone and dark-tone actors, with accuracy 87.5%, precision 75.0%, recall 100%, and F1-score 85.71%.

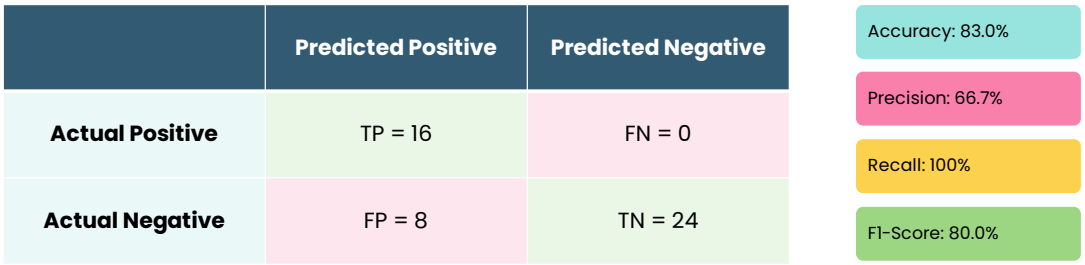


Figure 8. Confusion Matrix — Outdoor Testing, Medium-Brown Skin

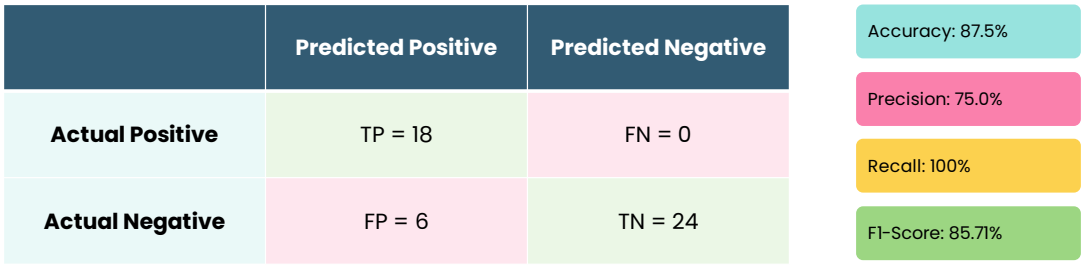


Figure 9. Confusion Matrix — Indoor Testing, Light Skin and Dark Skin

4.3. Comparison with Previous Studies

Table 6 situates the proposed system within the landscape of related sign language recognition studies. With a peak accuracy of 87.5%, the proposed approach matches or surpasses most previous BISINDO-specific implementations. The primary differentiating factor is the deliberate multi-group skin tone evaluation, which is absent from the majority of prior BISINDO studies that restrict participant diversity [3][4].

Table 6. Comparison with Previous Sign Language Recognition Studies

Study	Method	Dataset	Letter Class	Best Accuracy
Proposed	Faster R-CNN ResNet-50 + Skin Color Class Strategy	BISINDO (8,000 orig / 46,916 aug)	8 (A-H)	87.5%
[47]	CNN-based gesture classifier	BISINDO dataset	26	82.56%
[48]	Faster R-CNN (baseline)	BISINDO (controlled)	26	85.0%
[5]	SSD MobileNet	BISINDO (homogeneous)	8	83.7%
[6]	VGG-16	SIBI	2 (M and N)	87%
[49]	YOLOv5	BISINDO (real-time)	26	99.27%

4.4. Discussion

4.4.1. Main Findings

The experimental results yield four primary findings. First, the system achieves a peak accuracy of 87.5% and F1-score of 85.71% for light-skinned and dark-skinned actors under indoor conditions. Second, indoor testing consistently outperforms outdoor testing across all skin tone groups. Third, the performance gap across skin tone categories is minimal — at most 4.5 percentage points — confirming that the class-definition strategy successfully embeds skin tone diversity without producing severe inter-group bias. Fourth, recall consistently reaches 100% across all conditions, demonstrating reliable detection of target gestures regardless of the user's skin tone [50].

4.4.2. Comparison with Prior Research

The results are broadly consistent with and extend upon prior findings. Earlier BISINDO-focused studies using Faster R-CNN baselines on homogeneous datasets report accuracy values in the range of 83–85% [3],

[4], which the proposed system matches and exceeds under indoor conditions. Compared to CNN-LSTM hybrid approaches [51], the proposed system trades dynamic gesture support for substantially lower deployment complexity, operating at 24 FPS on standard hardware. Studies in computer vision fairness have highlighted that models trained on homogeneous datasets exhibit reduced generalization for underrepresented skin tones [52][53], the present study empirically demonstrates that this gap can be contained to within 4.5 percentage points through inclusive dataset construction and class-level encoding.

4.4.3. Implications of Findings

The results demonstrate that region-based detection architectures are viable for real-time BISINDO recognition when configured with skin-tone-aware class structures and trained on demographically diverse datasets. The consistent 100% recall across all groups indicates the system is well-suited for assistive communication contexts where missed detections are more disruptive than occasional false positives. The slight performance reduction observed for medium-brown skin tones under outdoor conditions suggests that future deployment environments requiring outdoor operation should incorporate illumination normalization preprocessing to mitigate reflectance-induced variability.

4.4.4. Strengths and Limitations

The primary strength of this study is its systematic evaluation of recognition fairness across skin tone diversity under heterogeneous environmental conditions, largely absent from prior BISINDO literature [3][4]. The primary limitations include: (1) coverage of only 8 of the 26 BISINDO alphabet classes; (2) the absence of a direct controlled comparison with YOLOv8 or SSD MobileNet on the same dataset; (3) the system has not been evaluated on dynamic backgrounds or with moving cameras; and (4) the participant pool of ten actors may not fully represent hand morphology variability in real-world deployment.

5. CONCLUSIONS

This study proposed a Faster R-CNN ResNet-50–based system for BISINDO alphabet recognition that explicitly incorporates skin color diversity through a 24-class definition strategy, representing eight alphabet gestures across three skin tone categories (light, medium-brown, and dark). The system was trained on a diverse dataset of 46,916 augmented images collected from ten actors under both indoor and outdoor conditions.

The best recognition performance was achieved under indoor testing conditions, with a peak accuracy of 87.5%, precision of 75.0%, recall of 100%, and F1-score of 85.71% for light-skinned and dark-skinned participants. Medium-brown-skinned participants achieved a maximum accuracy of 83.0%. The performance gap across skin tone categories was contained to within 4.5 percentage points, confirming that the system reliably detects target gestures regardless of the user's skin tone.

The theoretical contribution of this study lies in demonstrating that skin tone fairness in vision-based sign language recognition can be achieved through dataset diversity and class-level skin tone encoding, without requiring explicit skin segmentation preprocessing. The system operates at 24 frames per second on standard laptop hardware, confirming its viability for real-time assistive communication support.

The primary limitations include coverage of only 8 alphabet classes, moderate performance reduction in outdoor conditions, restriction to static gestures, and the absence of a direct controlled comparison with single-stage detectors. The system has not been evaluated on dynamic backgrounds or with moving cameras.

Future research directions include:

- Extending gesture class coverage to the full 26-letter BISINDO alphabet and dynamic gesture sequences.
- Conducting a controlled comparative evaluation between Faster R-CNN and YOLOv8/SSD MobileNet on the identical BISINDO dataset.
- Incorporating illumination normalization preprocessing to improve outdoor robustness.
- Expanding the participant pool across broader age groups, geographic regions, and hand morphologies.
- Investigating model compression techniques (knowledge distillation, quantization) to enable deployment on mobile and embedded platforms for broader accessibility among deaf communities.

DECLARATION

Supplementary Materials

Representative BISINDO alphabet images from different skin color categories, additional qualitative detection results, and detailed Faster R-CNN training configurations are provided to support reproducibility and further analysis.

Author Contribution
All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Sustainable Development Goals
This study supports **SDG 4 (Quality Education)** by enabling assistive technologies for Indonesian Sign Language learning and communication. It also contributes to **SDG 10 (Reduced Inequalities)** by addressing skin color diversity to promote fairness in vision-based recognition systems. Furthermore, the proposed Faster R-CNN–based approach aligns with **SDG 9 (Industry, Innovation, and Infrastructure)** through the development of robust and real-time AI solutions for inclusive human–computer interaction.

Funding
This research received no external funding.

Acknowledgement
The authors would like to thank all participants involved in the data collection process and Universitas Muslim Indonesia for the facilities and support provided during this research.

Conflicts of Interest
The authors declare no conflict of interest.

REFERENCES
[1] F. I. Pranto *et al.*, "A comprehensive image dataset of American Sign Language hand gestures," *Data Br.*, vol. 65, p. 112492, 2026, https://doi.org/10.1016/j.dib.2026.112492.
[2] I. M. M. Violet and R. L. Sri, "A comprehensive survey on recent advances and challenges in sign language recognition systems," *Discov. Artif. Intell.*, vol. 7, 2025, https://doi.org/10.1007/s44163-025-00629-7.
[3] L. N. Hayati, A. N. Handayani, W. Sakti, G. Irianto, and R. A. Asmara, "Improving Indonesian Sign Alphabet Recognition for Assistive Learning Robots Using Gamma-Corrected MobileNetV2," *Bul. Ilm. Sarj. Tek. Elektro*, vol. 7, no. 3, pp. 350–361, 2025, https://doi.org/10.12928/biste.v7i3.13300.
[4] S. Daniels, N. Suciati, and C. Fathichah, "Indonesian sign language recognition using yolo method," In *IOP Conference Series: Materials Science and Engineering*, vol. 1077, no. 1, p. 012029, 2021, https://doi.org/10.1088/1757-899X/1077/1/012029.
[5] L. N. Hayati, A. N. Handayani, W. S. G. Irianto, R. A. Asmara, D. Indra, and M. Fahmi, M. , "Classifying BISINDO Alphabet using Tensorflow Object Detection API," *Ilk. J. Ilm.*, vol. 15, no. 2, pp. 358–364, 2023, https://doi.org/10.33096/ilkom.v15i2.1692.358-364.
[6] N. H. Amir, C. Kusuma, and A. Luthfi, "Refining the Performance of Neural Networks with Simple Architectures for Indonesian Sign Language System (SIBI) Letter Recognition Using Keypoint Detection," *Ilk. J. Ilm.*, vol. 17, no. 1, pp. 64–73, 2025, https://doi.org/10.33096/ilkom.v17i1.2522.64-73.
[7] A. Umar, W. Qingshang, M. Badreddine, and Z. Jiangtao, "SignViT: An enhanced vision transformer framework for Attention-Based sign language hand gesture recognition," *Elsevier Biomed. Signal Process. Control*, vol. 112, no. 108602, 2026, https://doi.org/10.1016/j.bspc.2025.108602.
[8] Q. Zhu, J. Li, F. Yuan, J. Fan, and Q. Gan, "A Chinese Continuous Sign Language Dataset Based on Complex Environments," *arXiv preprint arXiv:2409.11960*, 2024, https://doi.org/10.48550/arXiv.2409.11960.
[9] S. T. Abd Al-Latief, S. Yussof, A. Ahmad, S. M. Khadim, and R. A. Abdulhasan, "Instant Sign Language Recognition by WAR Strategy Algorithm Based Tuned Machine Learning," *Int. J. Networked Distrib. Comput.*, vol. 12, no. 2, pp. 344–361, 2024, https://doi.org/10.1007/s44227-024-00039-8.
[10] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10023-10033, 2020, https://doi.org/10.1109/CVPR42600.2020.01004.
[11] Y. Yusmanto, H. Ar, and A. Prasetya, "Design and development of face recognition-based security system using expression game as liveness detection," *Bull. Soc. Informatics Theory Appl.*, vol. 8, no. 2, pp. 280–294, 2024, https://doi.org/10.31763/businta.v8i2.756.
[12] Y. Yang, T. Wan, M. Zhang, and L. Li, "High-precision YOLOv5 object detection technology based on multi-module collaborative optimization," *Multimed. Tools Appl.*, vol. 85, no. 2, p. 69, 2026, https://doi.org/10.1007/s11042-026-21227-4.
[13] J. Moni, R. R. Varghese, A. Binoy, B. S. Benny, L. Rajan and B. Benni, "Sign Language Translation Assistant using Machine Learning," *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, pp. 1062-1066, 2022, https://doi.org/10.1109/DASA54658.2022.9765097.

[14] N. Harlinda; Rendi, Ahmad; Azis, Huzain; Indra, Dolly; Hayati, Lilis Nur; Kurniati, "Classification of Cia-Cia Letters Using MobileNetV2 and CNN Methods," *2025 19th Int. Conf. Ubiquitous Inf. Manag. Commun.*, 2025, https://doi.org/10.1109/IMCOM64595.2025.10857478.
[15] J. Ahmed and B. Shawon, "A comparative analysis of video vision transformers on word-level sign language datasets," *PLoS One*, vol. 2, no. 21, pp. 1–22, 2026, https://doi.org/10.1371/journal.pone.0341909.
[16] M. Martawidjaja, R. D. Tandiono and L. Ayu Wulandhari, "Artificial Intelligence-Based System for Word Recognition and Extraction in Indonesian Sign Language Videos," *2025 8th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 802-808, 2025, https://doi.org/10.1109/ISRITI68345.2025.11393297.
[17] Y. V. Via, W. S. Saputra, M. I. Fachrurrozi, E. Y. Puspaningrum, F. T. Anggraeny, and S. R. Nudin, S. R. (2023)., "Object Localization and Detecting Alphabet in Sign Language BISINDO Using Convolution Neural Network," *Technium*, vol. 16, no. 143, 2023, https://doi.org/10.47577/technium.v16i.9973.
[18] F. M. Najib, "Sign language interpretation using machine learning and artificial intelligence," *Neural Comput. Appl.*, vol. 9, 2024, https://doi.org/10.1007/s00521-024-10395-9.
[19] A. M. J. AL Moustafa *et al.*, "Arabic Sign Language Recognition Systems: a Systematic Review," *Indian J. Comput. Sci. Eng.*, vol. 15, no. 1, pp. 1–18, 2024, https://doi.org/10.21817/indjcse/2023/v15i1/241501008.
[20] S. Al Ahmadi, F. Mohammad, and H. Al Dawsari, "Efficient YOLO-Based Deep Learning Model for Arabic Sign Language Recognition," *J. Disabil. Res.*, vol. 3, no. 4, 2024, https://doi.org/10.57197/jdr-2024-0051.
[21] H. Yang *et al.*, "Systemic representational biases in anatomical education: A multi-modal content analysis of racial, sex, and skin tone diversity across literature, textbooks, and digital platforms," *Ann. Anat. - Anat. Anzeiger*, vol. 265, 2026, https://doi.org/10.1016/j.aanat.2026.152801.
[22] S. M. K. Raja'a M. Mohammed, "Iraqi Sign Language Translator system using Deep Learning," *Al-Salam J. Eng. Technol.*, vol. 1, pp. 109–116, 2023, https://doi.org/10.55145/ajest.2023.01.01.0013.
[23] S. Aiouez, A. Hamitouche, M. Belmadoui, K. Belattar, and F. Souami, "Real-time Arabic Sign Language Recognition based on YOLOv5," *SCITEPRESS*, vol. 17–25, pp. 17–25, 2022, https://doi.org/10.5220/0010979300003209.
[24] W.-B. Ma, Y. Yang, and W.-C. Fang, "An Effective Tuberculosis Detection System Based on Improved Faster R-CNN with RoI Align Method," *2023 IEEE Biomed. Circuits Syst. Conf.*, 2023, https://doi.org/10.1109/BioCAS58349.2023.10388704.
[25] M. R. Ningsih *et al.*, "Sign Language Detection System Using YOLOv5 Algorithm to Promote Communication Equality People with Disabilities," *Sci. J. Informatics*, vol. 11, no. 2, pp. 549–558, 2024, https://doi.org/10.15294/sji.v11i2.6007.
[26] G. Priyadharshini and D. R. Judie Dolly, "Comparative Investigations on Tomato Leaf Disease Detection and Classification Using CNN, R-CNN, Fast R-CNN and Faster R-CNN," *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 1540-1545, 2023, https://doi.org/10.1109/ICACCS57279.2023.10112860.
[27] I. Panopoulos, E. Topalis, and N. Petrellis, "Greek Sign Language Detection with Artificial Intelligence," *Electronics*, vol. 14, no. 16, p. 3241, 2025, https://doi.org/10.3390/electronics14163241.
[28] Y. Farhan, Z. Haimer, and A. A. Madi, "Real-time interpretation of American Sign Language using SSD-MobileNet," *International Journal of Computational Vision and Robotics*, vol. 16, no. 2, pp. 222-245. 2026, https://doi.org/10.1504/IJCVR.2026.151535.
[29] B. Tej, F. Nasri, and A. Mtibaa, "Detection of Pepper and Tomato leaf diseases using deep learning techniques," *Proc. 2022 5th Int. Conf. Adv. Syst. Emergent Technol. IC_ASET 2022*, no. January 2024, pp. 149–154, 2022, https://doi.org/10.1109/IC_ASET53395.2022.9765923.
[30] A. S. G. Raharjo and E. Sugiharti, "Alphabet Classification of Sign System Using Convolutional Neural Network with Contrast Limited Adaptive Histogram Equalization and Canny Edge Detection," *Sci. J. Informatics*, vol. 10, no. 3, pp. 239–250, 2023, https://doi.org/10.15294/sji.v10i3.44137.
[31] S. Feng, L. Zhao, H. Shi, M. Wang, S. Shen, and W. Wang, "One-dimensional VGGNet for high-dimensional data," *Appl. Soft Comput.*, vol. 135, p. 110035, 2023, https://doi.org/10.1016/j.asoc.2023.110035.
[32] S. Wang, "Domain-adaptive faster R-CNN for non-PPE identification on construction sites from body-worn and general images," *Sci. Rep.*, pp. 1–22, 2026, https://doi.org/10.1038/s41598-026-35148-7.
[33] I. Lamaakal, C. Yahyati, Y. Maleh, and K. El Makkaoui, "An explainable hybrid CNN – transformer model for sign language recognition on edge devices using adaptive fusion and knowledge distillation," *Sci. Rep.*, vol. 16, no. 7143, pp. 1–23, 2026, https://doi.org/10.1038/s41598-026-38478-8.
[34] V. E. Nurfita Sari, A. Arifin and F. Arrofiqi, "Design of Two-Way Indonesian Sign Language System based on Smart-Glove with Artificial Neural Network Classification Method," *2024 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, pp. 1-6, 2024, https://doi.org/10.1109/CENIM64038.2024.10882818.
[35] P. Hu, J. Lu, Y. Cui, B. Hu, and F. Liang, "Insulator Defect Detection based on Faster R-CNN and YOLOv3 Algorithm," *2023 IEEE 4th China Int. Youth Conf. Electr. Eng.*, 2023, https://doi.org/10.1109/CIYCEE59789.2023.10401677.
[36] N. Azizah, "Improving Smear-Negative Tuberculosis Detection Using Data Augmentation and Faster R-CNN," *Int. J. Cyber IT Serv. Manag.*, vol. 6, no. 1, pp. 65–77, 2026, https://doi.org/10.34306/ijcitsm.v6i1.233.

[37] H. M. Hamza and A. Wali, "Pakistan Sign Language Recognition: From Videos to Images," *Signal, Image Video Process.*, vol. 19, no. 8, p. 682, 2025, https://doi.org/10.1007/s11760-025-04230-4.
[38] D. A. Rachmawati, C. Supriyanto, M. A. Soeleman and N. Hendriyanto, "Performance Improvement of the Faster R-CNN Model Using ResNet Architecture and Data Augmentation for Indonesian Sign Language (BISINDO) Detection," *2025 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 298-304, 2025, https://doi.org/10.1109/ISemantic67418.2025.11291950.
[39] Y. Tian, Y. Dong, M. Ahmed, S. O. Shah, and E. Alabdulkreem, "Real-Time Chinese Sign Language Recognition Based on Convolutional Neural Network," *Int. J. Humanoid Robot.*, p. 2540022, 2026, https://doi.org/10.1142/S0219843625400225.
[40] W. Luo, "Research on gesture recognition based on YOLOv5," *2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pp. 447-450, 2022, https://doi.org/10.1109/ICBAIE56435.2022.9985931.
[41] D. A. Rachmawati, C. Supriyanto, M. A. Soeleman and N. Hendriyanto, "Performance Improvement of the Faster R-CNN Model Using ResNet Architecture and Data Augmentation for Indonesian Sign Language (BISINDO) Detection," *2025 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 298-304, 2025, https://doi.org/10.1109/ISemantic67418.2025.11291950.
[42] E. L. R. Ewe, C. P. Lee, L. C. Kwek, and K. M. Lim, "Hand Gesture Recognition via Lightweight VGG16 and Ensemble Classifier," *Appl. Sci.*, vol. 12, no. 15, 2022, https://doi.org/10.3390/app12157643.
[43] M. G. F. Odounfa, C. D. S. J. Gbemavo, S. P. G. Tahi, and R. L. Glèlè Kakaï, "Deep learning methods for enhanced stress and pest management in market garden crops: A comprehensive analysis," *Smart Agric. Technol.*, vol. 9, p. 100521, 2024, https://doi.org/10.1016/j.atech.2024.100521.
[44] A. Alayed, "Machine Learning and Deep Learning Approaches for Arabic Sign Language Recognition: A Decade Systematic Literature Review," *Sensors*, vol. 24, no. 23, 2024, https://doi.org/10.3390/s24237798.
[45] Z. Haimer, K. Mateur, Y. Farhan and A. A. Madi, "Road Marking Detection and Instance Segmentation Using YOLOv8 Models," *2024 10th International Conference on Optimization and Applications (ICOA)*, pp. 1-8, 2024, https://doi.org/10.1109/ICOA62581.2024.10754303.
[46] D. O. Pratama, U. N. Malang, A. N. Handayani, and U. N. Malang, "Development of Embedded System Learning Module Using Project-based Learning Method for Industrial Electronics Department," *Lect. J. Pendidik.*, vol. 16, pp. 225–238, 2025, https://doi.org/10.31849/lectura.v16i1.25415.
[47] N. Ahmad, E. S. Wijaya, C. Tjoaquinn, H. Lucky, and I. A. Iswanto, "Transforming Sign Language using CNN Approach based on BISINDO Dataset," *2023 Int. Conf. Informatics, Multimedia, Cyber Informations Syst.*, 2023, https://doi.org/10.1109/ICIMCIS60089.2023.10349011.
[48] D. Joan, V. Vincent, K. J. Daniel, S. Achmad and R. Sutoyo, "BISINDO Hand-Sign Detection Using Transfer Learning," *2023 IEEE 8th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pp. 1-7, 2023, https://doi.org/10.1109/ICRAIE59459.2023.10468194.
[49] M. Z. U. Rahman *et al.*, "A Real-Time System for Classification of Pakistan Sign Language using Machine Learning," *2023 17th International Conference on Open Source Systems and Technologies (ICOSST)*, pp. 1-5, 2023, https://doi.org/10.1109/ICOSST60641.2023.10414200.
[50] A. Youssef, A. Gaber, and S. M. EL-Metwally, "A Dual-Architecture Deep Learning Pipeline for Real- Time High-Accuracy Arabic Sign Language Recognition," *Discov. Artif. Intell.*, 2026, https://doi.org/10.21203/rs.3.rs-8605046/v1.
[51] R. Dababo and M. -H. Wang, "Speech2Sign with ASL-Transformers and SL-GAN," *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, pp. 965-971, 2024, https://doi.org/10.1109/ICICML63543.2024.10958134.
[52] K. Malik, C. Robertson, S. A. Roberts, T. K. Remmel, and J. A. Long, "Computer vision models for comparing spatial patterns: understanding spatial scale," *Int. J. Geogr. Inf. Sci.*, vol. 37, no. 1, pp. 1–35, 2023, https://doi.org/10.1080/13658816.2022.2103562.
[53] R. A. Asmara *et al.*, "YOLO-based object detection performance evaluation for automatic target aimbot in first-person shooter games," *Bull. Electr. Eng. Informatics*, vol. 13, no. 4, pp. 2456–2470, 2024, https://doi.org/10.11591/eei.v13i4.6895.

AUTHOR BIOGRAPHY

Lilis Nur Hayati, master's degree in Information Technology was obtained in 2005 from Universitas Gadjah Mada, and the Doctoral degree in the same field is currently being pursued at Universitas Negeri Malang. Currently, she is a lecturer in Information Systems with 9 years of teaching experience at Universitas Muslim Indonesia. Her research interests include software design, software requirements analysis, decision support systems, technopreneurship, human-computer interaction, e-business concepts, research methodology, and operating systems.
Email:lilis.nurhayati.2205349@students.um.ac.id
Google Scholar https://scholar.google.com/citations?hl=id&user=me_9y28AAAAJ

Anik Nur Handayani, master's degree in Electrical Engineering in 2008 from Institut Teknologi Sepuluh Nopember (ITS) Surabaya, Indonesia, and earned her Doctoral degree in Science and Advanced Engineering from Saga University, Japan.

She is currently a university lecturer at Universitas Negeri Malang, Indonesia. Her research interests include image processing, biomedical signal analysis, artificial intelligence, machine learning, deep learning, computer vision, and assistive technologies.
Email: aniknur.ft@um.ac.id
Google Scholar:
https://scholar.google.com/citations?hl=en&user=nqPHjbMAAAAJ&view_op=list_works&sortby=pubdate

Wahyu Sakti Gunawan Irianto, master's degree in Computer Science from Universitas Indonesia, Jakarta, in 1997. He earned a Doctoral degree in Computer Science (M.Kom). He is currently a senior lecturer in the Department of Electrical Engineering at Universitas Negeri Malang, Indonesia. His research interests include computer science education, educational technology, intelligent systems, embedded and microcontroller applications, and digital systems. He has contributed to various projects, such as an interactive learning module based on Arduino and multimodal dataset research in the LUMINA project.
Email: wahyu.sakti.ft@um.ac.id
Google Scholar: https://scholar.google.com/citations?user=DAWTUlAAAAAJ&hl=en

Rosa Andrie Asmara, received his Bachelor's degree in Electronics Engineering from Universitas Brawijaya, Malang, in 2004. He obtained his Master's degree in Computer Science from Institut Teknologi Sepuluh Nopember, Surabaya, in 2009, and completed his Doctoral degree in Computer Science at Saga University, Japan, in 2013. He is currently a lecturer at Politeknik Negeri Malang, Indonesia. His research interests include machine learning, image understanding, and computer vision.
Email: rosa_andrie@polinema.ac.id
Google Scholar: https://scholar.google.co.id/citations?user=A1592kEAAAAJ&hl=en

Dolly Indra, earned his Doctoral degree in Information Technology from Universitas Gunadarma in 2017. He is currently a lecturer at the Faculty of Computer Science, Universitas Muslim Indonesia. His research interests include image processing, computer vision, microcontroller systems, and information systems.
Email: dolly.indra@umi.ac.id
Google Scholar: https://scholar.google.co.id/citations?user=94_nu_QAAAAJ&hl=en

Nor Salwa Damanhuri, received her Bachelor of Science (Hons.) in Electrical and Electronics Engineering from Universiti Tenaga Nasional (UNITEN), Malaysia, in March 2002. She completed her Master of Science in Control Systems Engineering at The University of Sheffield, United Kingdom, in September 2005, and obtained her Doctor of Philosophy (Ph.D.) in Bioengineering from the University of Canterbury, New Zealand, in April 2015. She is currently an Associate Professor at the Centre for Electrical Engineering Studies, Universiti Teknologi MARA (UiTM), Penang Branch, Malaysia. Her research interests include biomedical engineering, digital signal processing, mathematical modeling, control systems, and solar PV system applications.
Email: norsalwa071@uitm.edu.my
Google Scholar: https://scholar.google.com/citations?user=O3DojDMAAAAJ&hl=en