ISSN 2301-8984 (Print), 2809-1884 (Online)
Published by Institute of Computer Science

Alphabet SIBI sign language recognition using YOLOv11 for real-time gesture detection

Salsabilla Azahra Putri[1], Murinto[2], Sunardi[3]
[1] Master of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[2] Department of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[3] Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

ARTICLE INFO

Article history:

Received May 28, 2025
Revised Jun 27, 2025
Accepted Jun 30, 2025

Keywords:

Backbone optimization;
Data augmentation;
Gesture recognition;
Real-time detection;
Sibi alphabet.

ABSTRACT

Modern gesture recognition systems for sign language face challenges in balancing computational efficiency and detection accuracy in complex and dynamic environments. To address this, this study proposes a SIBI alphabet recognition framework based on YOLOv11, optimized for real-time applications. The model architecture integrates a modified, efficient YOLOv11 backbone to enable accurate hand gesture feature extraction with minimal latency. A custom SIBI dataset comprising alphabet signs and essential vocabulary is used to train the model, supported by data augmentation techniques that enhance robustness against variations in position, lighting, and background. Experimental results demonstrate that the model achieves a high detection accuracy with an mAP50 of 97%, while significantly reducing computational complexity. These findings present a meaningful scientific contribution by showcasing how a lightweight yet highly accurate deep learning model can be effectively applied to sign language recognition, particularly for SIBI in the Indonesian context. From a practical standpoint, this framework offers a real-time gesture detection solution that is suitable for deployment on resource-constrained devices, making it accessible for mobile or embedded systems. The system can replace or complement traditional communication aids, especially in inclusive education, public services, and healthcare. Furthermore, the proposed method can be adapted for gesture-based interaction in other domains such as athletic training, physical education, and app-based fitness programs where accurate and real-time motion recognition is essential.

Corresponding Author:

Salsabilla Azahra Putri
Master of Informatics ,
Universitas Ahmad Dahlan ,
Yogyakarta, Indonesia, 55191
Email: salsabillaazahra06@gmail.com

1. INTRODUCTION

Language is the primary medium of human communication. However, for individuals with hearing impairments, limitations in using spoken language lead them to rely on sign language as the main means of conveying information and interacting with their social environment. In Indonesia, the Indonesian Sign Language System (SIBI) is the official language used by the deaf community and has been standardized based on Indonesian grammar (Safitri, Yuniarno, & Rachmadi, 2024). Unfortunately, public understanding of SIBI remains very limited. This situation creates a communication gap between deaf individuals and the general public who do not understand sign language, affecting social inclusion, education, and access to public services.

According to data from Indonesia's National Socio-Economic Survey (Susenas) and the Ministry of Social Affairs, there are over 470,000 individuals with hearing disabilities across the country as of 2022. However, due to stigma and limited public awareness, many of them still experience marginalization in accessing essential services such as education, health care, and employment. The development of automated sign language recognition systems is therefore critical to enhancing social inclusion, digital accessibility, and equitable communication, particularly in public service domains. Such technologies can serve as assistive tools that empower the deaf community and bridge communication gaps in multilingual and multicultural environments like Indonesia (Affairs, 2023; Statistik, 2022)

With the advancement of technology particularly in the fields of Artificial Intelligence (AI) and Computer Vision research on automatic sign language recognition has seen significant progress. These technologies aim to detect and recognize hand movements or gestures that represent sign language symbols, which are then translated into text or speech that can be understood by the general public (Ong & Ranganath, 2005). This approach has the potential to be an inclusive solution that bridges the communication gap between people with hearing disabilities and the wider community.

Various methods have been proposed for sign language recognition, ranging from wearable sensor-based approaches to systems based on visual image processing. Sensor-based approaches, such as those using accelerometers and flexible sensors, offer advantages in measurement accuracy, but are less suitable for everyday use due to their intrusive nature and high cost (Sümbül, 2024). In contrast, computer vision-based approaches that utilize cameras to observe and process hand movements are considered more practical and hold great potential for real-world applications, as they do not require the user to wear additional equipment (Shivaprasad Yadav, Itagi, Krishna Suresh, K.L, & A C, 2023).

In the domain of computer vision, deep learning algorithms have been widely employed for gesture recognition. Popular methods include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and pose estimation-based models (Selvaraju et al., 2020)(Guo et al., 2022). CNNs have proven effective in extracting spatial features from images, while RNNs are capable of modeling temporal dependencies in sequential movements found in dynamic gestures. However, these approaches often require multi-stage processing and may not be suitable for real-time implementation, particularly in applications demanding fast and efficient detection.

As a more efficient alternative, one-stage object detection algorithms such as You Only Look Once (YOLO) offer high speed and accuracy by detecting objects directly from input images. YOLO integrates both detection and classification processes into a single inference step, making it highly suitable for real-time applications like sign language recognition (He, Zhou, Liu, & Ma, 2024). Since its initial introduction by Redmon et al., the YOLO algorithm has undergone significant improvements in architectural design, small object detection, and computational efficiency. Its latest iteration, YOLOv11, offers performance enhancements over previous versions (B. Liu & Li, 2024).

YOLOv11 introduces several new innovations, including a deeper yet lightweight backbone structure, more adaptive label assignment strategies, and stronger generalization capabilities for small or background-complex objects (Zanevych et al., 2024). These features are crucial in hand gesture recognition, where variations in position, orientation, and lighting frequently occur in real-world environments. Furthermore, YOLOv11's ability to perform high-speed inference enables its direct use in interactive applications powered by webcams or mobile devices.

Although many studies have developed sign language recognition systems using the YOLO approach, most remain limited to international sign languages such as American Sign Language (ASL) (Y. Liu, Nand, Hossain, Nguyen, & Yan, 2023), British Sign Language (BSL), or other non-standard gesture systems. In contrast, research on Indonesian Sign Language (SIBI), particularly in its static alphabet form (A–Z), is still relatively scarce in the scientific literature. Yet, the SIBI alphabet forms the fundamental building block for word and sentence construction, making its accurate recognition a crucial first step for developing more complex sign language translation systems.

Although there have been a few initial efforts to build SIBI alphabet recognition systems in Indonesia, most remain at a prototyping stage with limited scalability and real-time performance. For example, research by (Safitri et al., 2024) utilized traditional hand segmentation and contour detection techniques for static alphabet classification, but the model was highly sensitive to lighting and background variations. These systems also lack the computational efficiency required for deployment on mobile or embedded platforms, and often do not support interactive use cases. As a result, practical applications of existing SIBI recognition systems remain restricted, particularly in public services, mobile communication, or inclusive educational tools.

This research aims to develop a real-time recognition system for the SIBI alphabet using the YOLOv11 algorithm. The system is designed to detect hand gesture alphabets accurately and efficiently, and to be seamlessly integrated with a real-time, camera-based interface. The implementation of YOLOv11 is expected to overcome common challenges in hand gesture recognition, such as dynamic backgrounds, lighting variations, and differences in hand size and position. By combining fast detection capabilities with an optimized architecture, this study contributes to the development of inclusive technological solutions that support communication between the deaf community and the general public in Indonesia.

Sign language recognition has become an increasingly important research domain in the context of inclusive technology, particularly for individuals with hearing disabilities. Various approaches have been developed to identify hand gestures that represent words or letters in sign languages. Early methods primarily relied on sensor-based devices such as data gloves and accelerometers. While these systems offered high accuracy, they were less practical due to the requirement for users to wear special equipment, which reduced comfort and flexibility for daily use (Dignan, Perez, Ahmad, Huber, & Clark, 2022).

With the advancement of computer vision and artificial intelligence technologies, camera-based and image-processing approaches have become more dominant in sign language recognition research. The application of Convolutional Neural Networks (CNNs) has significantly improved accuracy in recognizing both static and dynamic gestures. CNNs are effective at extracting spatial features from hand images with high precision. However, their real-time performance remains limited, especially when deployed on resource-constrained devices (Tourani, Soroori, Shahbahrami, & Akoushideh, 2021).

Recent studies have begun integrating real-time object detection methods to address these challenges. Among them is the YOLO (You Only Look Once) algorithm, known for its efficiency and speed in identifying objects in a single-stage process. YOLOv5 and YOLOv7 have been widely applied in recognizing sign language alphabets using datasets such as American Sign Language (ASL), yielding promising results in terms of both precision and inference speed (Gomez & Arzuaga, 2024). These studies confirm that one-stage detection approaches like YOLO are well-suited for real-time systems, particularly in applications such as automatic sign language translation.

Nevertheless, research on sign language recognition in local contexts such as Indonesian Sign Language (SIBI) remains very limited. Most studies in Indonesia have focused on static image classification using CNN or MobileNet, without addressing object detection in real-time or complex background conditions (Agrawal, Kumar, Kumar, & Arvind, 2024). Additional challenges include the lack of locally available datasets, limited diversity in viewpoints, lighting conditions, and individual variations.

With the introduction of YOLOv11, the latest version of the YOLO algorithm, offering enhanced computational efficiency and improved small object detection accuracy, its application in recognizing SIBI hand gestures is increasingly relevant. YOLOv11 features a more lightweight yet effective backbone architecture, as well as updates to its loss function and training process, enabling high-precision detection with low latency even in uncontrolled environments (Chen et al., 2022).

To date, few studies have specifically explored the use of YOLOv11 for real-time SIBI alphabet recognition. Therefore, this research aims to provide an initial and significant contribution by combining state-of-the-art object detection models with local needs, while supporting the development of inclusive and applicable assistive technologies in Indonesia.

2. Methods

This study aims to develop a real-time detection system for the Indonesian Sign Language (SIBI) alphabet using the YOLOv11 algorithm. The methodological stages include data collection, data preprocessing, model architecture design, model training, performance evaluation, and real-time implementation.

2.1. Dataset and Preprocessing

To develop an efficient and accurate YOLOv11-based recognition system for the Indonesian Sign System Alphabet (SIBI), a high-quality training dataset is essential. The dataset must be representative of the various hand gesture variations used in the SIBI alphabet, including diversity in hand positions, backgrounds, lighting conditions, and inter-individual gesture variations.

In this study, the dataset was collected through high-resolution image acquisition using a digital camera under both natural and artificial lighting conditions. Each SIBI alphabet gesture was captured from multiple angles and with different background settings to ensure the model's ability to generalize across real-world environments. A total of 26 SIBI alphabet letters were collected, with each letter performed by several different subjects to enrich the data distribution. An overview of the dataset distribution is illustrated in Figure 1.

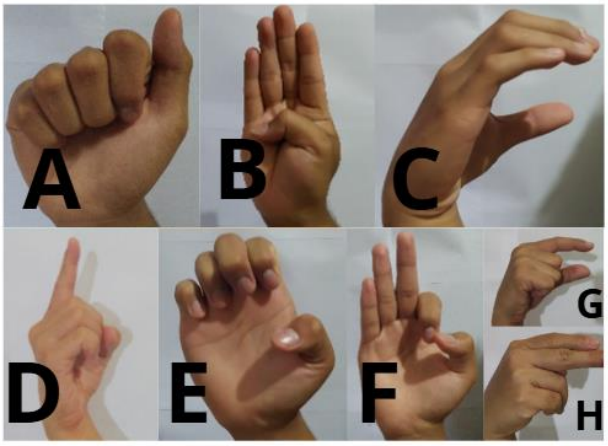

Figure 1. Dataset SIBI Language

Following image acquisition, the dataset underwent a data cleaning process to eliminate blurry, overexposed, or unclear gesture images. Only images with clearly represented hand gestures were retained for annotation. Annotation was performed using LabelImg, with labels saved in YOLOv11 format (.txt), which specifies the bounding box coordinates and class labels (A–Z) for each hand gesture.

To ensure the system performs effectively under diverse conditions, data augmentation techniques were applied. These techniques included rotation (±15°), scaling, horizontal flipping, contrast adjustments, and artificial lighting variations to simulate low-light environments. Additionally, color normalization was applied to improve model robustness against variations in skin tone and lighting. The primary visual features targeted for detection in this dataset are finger shapes and hand positions, which distinguish one letter from another in the SIBI alphabet. Since several letters have similar hand shapes (e.g., letters *N* and *M*), preserving spatial resolution and local features such as finger contours was critical during preprocessing. Therefore, all images were resized to YOLOv11's standard input dimension (640×640 px) while maintaining the original aspect ratio to avoid visual distortion.

A total of 2,079 images were initially collected, covering all 26 SIBI alphabet letters. Each letter was represented by approximately 80 images, performed by multiple individuals with varying backgrounds and lighting conditions to enhance dataset diversity. Following preprocessing and augmentation including rotation, flipping, contrast adjustment, and color normalization the dataset expanded to 5,405 labeled images.

The dataset was then divided using stratified sampling into three subsets: 80% for training (4,324 images), 10% for validation (541 images), and 10% for testing (540 images). This stratification ensured balanced representation of each alphabet class across all subsets.

2.2. Experimental Setup

To ensure reproducibility and transparency, the model training and evaluation were conducted on a system with the following hardware and software specifications in Table 1:

Table 1. Experimental Setup Summary

Category	Specification
Proccessor	Intel Core i7-12700H @ 2.3GHz
RAM	8 GB DDR4
Operating System	Windows 11
GPU	Intel Iris Xe (Integrated GPU)

The training process utilized the following hyperparameters in Tabel 2:

Table 2. Training Settings

Training Setting	Details
Input Image Size	640x640 pixels
Batch Size	32
Epoch	100
Validation Split	80% traininng, 10% validation, 10% testing

2.3. YOLOv11

YOLO (You Only Look Once) is a real-time object detection algorithm that has consistently demonstrated superior performance in terms of speed, efficiency, and accuracy across a wide range of object detection tasks (Sani, Zolfagharian, & Kouzani, 2024)(Aboud, Elsayaad, Gad, Abdelaziz, & Atia, 2024)(Ali & Zhang, 2024). Its latest version, YOLOv11, features a modular architecture comprising four main components: the input layer, backbone, neck, and detection head (Sani et al., 2024)(He et al., 2024).

This architecture is specifically designed to support real-time detection applications, including SIBI alphabet recognition, which requires fast and accurate classification of various hand shapes and positions. YOLOv11's improvements make it especially suitable for capturing subtle variations in static hand gestures under diverse environmental conditions.

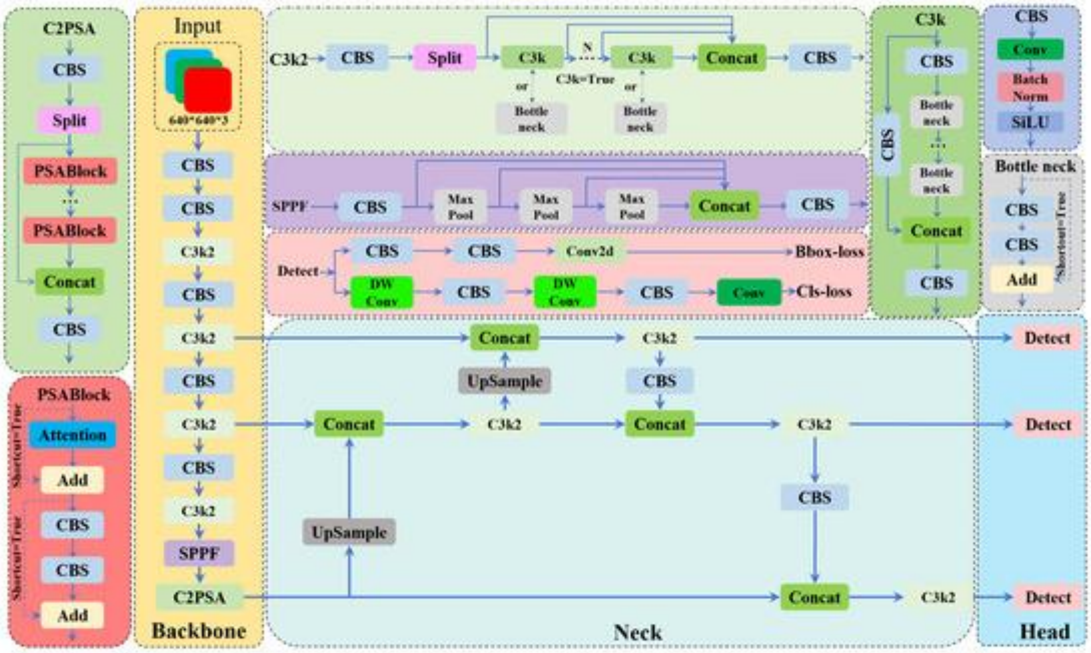

Figure 2. Architecture of YOLOv11

As illustrated in Figure 2, the backbone of YOLOv11 employs the C3k2 block (Cross-Stage Partial block with kernel size 2), which consists of two compact convolutional layers designed to efficiently extract features from input images (Das, Sayma, Barman, & Hasan, 2024). This block enhances the spatial and structural representation of hand gestures, which is essential for distinguishing between letters with similar finger shapes.

To support multi-scale detection, YOLOv11 integrates Spatial Pyramid Pooling-Fast (SPPF), enabling the fusion of spatial information from multiple resolutions (Zhang et al., 2024). This improves the model's robustness in detecting gestures under diverse scales and orientations.

In addition, the C2PSA block (Convolutional Block with Parallel Spatial Attention) is implemented to enhance the model's attention toward critical regions in the image, such as fingertips and palm contours that form sign gestures. By employing parallel spatial attention mechanisms, this block improves detection accuracy for small-scale and variable-position targets commonly encountered in sign language data (B. Liu & Li, 2024).

The neck structure in YOLOv11 incorporates feature pyramids using C3k2 blocks combined with dual-path feature aggregation (Ahmed, Farhan, Eesaar, Chong, & Tayara, 2024). During the upsampling phase, the CBS block is used to refine features, while during downsampling, cross-stage connections fuse shallow and deep features to strengthen the semantic representation of gesture information.

To reduce computational complexity without sacrificing accuracy, the YOLOv11 detection head employs dual depthwise convolution (DWConv) layers (Rodríguez-Lira et al., 2024). These layers significantly reduce the number of parameters and floating-point operations (FLOPs), making YOLOv11 highly efficient and suitable for edge devices or mobile platforms (Yang et al., 2024).

YOLOv11 is available in various model scales n, s, m, l, x which share similar architectures but differ in parameter counts. This makes it adaptable to both small- and large-scale object detection scenarios. In the context of SIBI alphabet recognition, lightweight variants such as YOLOv11n or YOLOv11s are ideal, as the system is designed to operate in real-time on resource-constrained devices like webcams or embedded systems (He et al., 2024).

To further optimize performance for sign language recognition, minor modifications were applied to the backbone and neck structures. These adjustments adapt the model to capture finer gesture features and enhance spatial precision. Such customization maintains low inference latency while improving accuracy in detecting complex hand positions and finger contours (Parico & Ahamed, 2021).

2.4. Performance Evaluation

To evaluate the performance of the enhanced model and validate the effectiveness of each architectural improvement, several key evaluation metrics were employed: Precision (P), Recall (R), Average Precision (AP), Mean Average Precision (mAP), F1-score, Number of Parameters (Params), Floating Point Operations (FLOPs), and Detection Time per Image. Among these metrics, Params and FLOPs are used to measure the model's computational complexity, while P, R, mAP, and F1-score are used to assess detection accuracy (Roy, Bhaduri, Kumar, & Raj, 2023).

In model evaluation, True Positives (TP) refer to the areas correctly predicted by the model as valid hand gestures, meaning the prediction matches the ground truth (Xiao & Kang, 2021). False Positives (FP) indicate regions where the model incorrectly identifies background elements as hand gestures i.e., predicting a gesture when in fact there is none. False Negatives (FN) refer to areas where the model fails to detect actual hand gestures i.e., when the model predicts background even though a gesture is present. In this study, the total number of detections made by the model consists of TP + FP, whereas TP + FN represents the total number of actual hand gestures present in the image.

In hand gesture detection, Precision (P) is a key metric to evaluate the model's ability to avoid false alarms. It calculates the proportion of correct hand gesture predictions among all gesture predictions made by the model, and it is defined by the following formula:

$$P = TP/(TP + FP) \qquad (1)$$

Recall (R) is a metric that measures the model's ability to correctly identify actual hand gestures among all gestures that are truly present in the dataset. A higher recall value indicates that the model is missing fewer detections, thus performing better in terms of capturing all relevant gestures. Recall is calculated using the following formula:

$$R = TP/(TP + FN) \qquad (2)$$

Average Precision (AP) is calculated as the area under the precision-recall (PR) curve. The detection threshold determines the trade-off between precision and recall. In single-class detection, AP integrates the precision and recall values across all confidence thresholds. It is mathematically defined by the following formula:

$$AP = \int_0^1 P(R)dR \qquad (3)$$

Here, P(R) represents the precision at a given recall level RRR. In multi-class object detection, mean Average Precision (mAP) is computed by averaging the AP values across all object categories. The formula is defined as follows:

$$mAP \ = \frac{1}{N}\sum_{1}^{N} APi \qquad (4)$$

Where APi denotes the Average Precision for the *i*-th category, and NNN represents the total number of categories. The F1-score is a crucial metric that evaluates the balance between precision and recall in hand gesture detection models. It provides a comprehensive performance assessment by harmonizing both metrics into a single value. The F1-score is computed as the harmonic mean of precision and recall, as defined by the following formula:

$$F1 - Score = 2PR/(P+R) \qquad (5)$$

In this study, a higher mean Average Precision (mAP) value indicates superior detection performance and improved recognition accuracy of the model. By optimizing these evaluation metrics, the model enhances its practical effectiveness, particularly in recognizing complex SIBI hand gestures in real-world environments.

3. Result

This study produced a YOLOv11-based detection model for recognizing SIBI alphabets, trained using a diverse dataset of hand gesture images. The evaluation was conducted both quantitatively, using performance metrics, and qualitatively, through visual analysis of detection results. The following discussion presents the findings based on the outcomes of the training and testing phases.

3.1. Training

Table 3 summarizes the training experimental environment, while Table 4 lists the key parameters used in the development of the YOLOv11-based SIBI alphabet recognition model. A custom dataset was created, consisting of 2,079 annotated hand gesture images representing the complete SIBI alphabet. This dataset was then split into training, validation, and testing sets with a ratio of 8:1:1.

Table 3. Exprrimental Conditions

Experimental Environment	*Details*
Programming Language	Python 3.8
Operating System	Windows 11
Deep Learning framework	Pytorch 2.0.1
CPU	Intel Core i7
GPU	Intel Core

Table 4. Key Parameters of Experimental Training

Traininng Parameters	*Details*
Epochs	100
Batch Size	16
Image Size	640x640
Optimazation algorithm	None
Pre-training weight	None

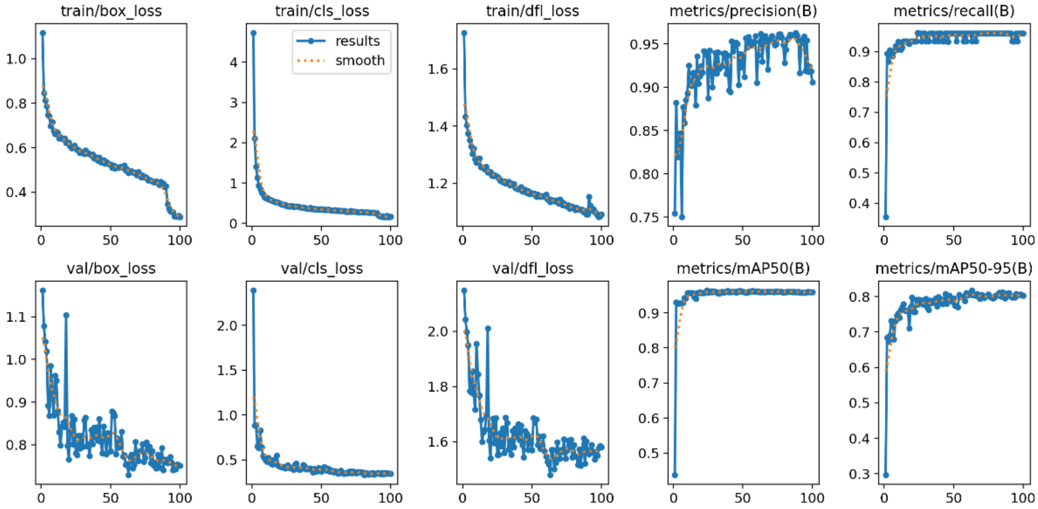

Figure 3. Result Traing

Figure 3 illustrates the training performance of the YOLOv11 model in real-time detection of SIBI alphabet sign language. Over the course of 100 training epochs, the loss values for both training and validation datasets showed a consistent downward trend. Box loss, classification loss (cls loss), and Distribution Focal Loss (DFL loss) on the training set decreased significantly, indicating that the model effectively learned spatial representations and gesture classifications. Box loss dropped from over 1.0 to below 0.3, while classification loss, initially above 4.0, stabilized around 0.2 by the end of training. Similarly, DFL loss also showed a steady decline, reflecting the model's improved ability to estimate object positions with high precision.

On the validation set, although loss values fluctuated slightly, the overall trend still showed a decrease, indicating that the model did not suffer from significant overfitting and retained its ability to generalize to unseen data. Additionally, evaluation metrics such as precision and recall demonstrated strong performance. Precision increased steadily to over 95%, indicating that most of the model's predictions were correct (i.e., minimal false positives). Recall also rose sharply from the early training stages and remained around 94%, suggesting the model was able to detect nearly all gestures in the dataset (i.e., minimal false negatives).

Furthermore, the model's overall performance in detecting SIBI alphabets is reflected in its mean Average Precision (mAP) scores. The mAP@0.5 reached approximately 97%, while mAP@0.5:0.95 remained above 80%, demonstrating the model's robustness in detecting gestures with varying degrees of difficulty and object overlap. These results reinforce the potential of YOLOv11 as an efficient and accurate solution for real-time SIBI alphabet recognition, particularly for deployment in inclusive communication systems powered by computer vision.

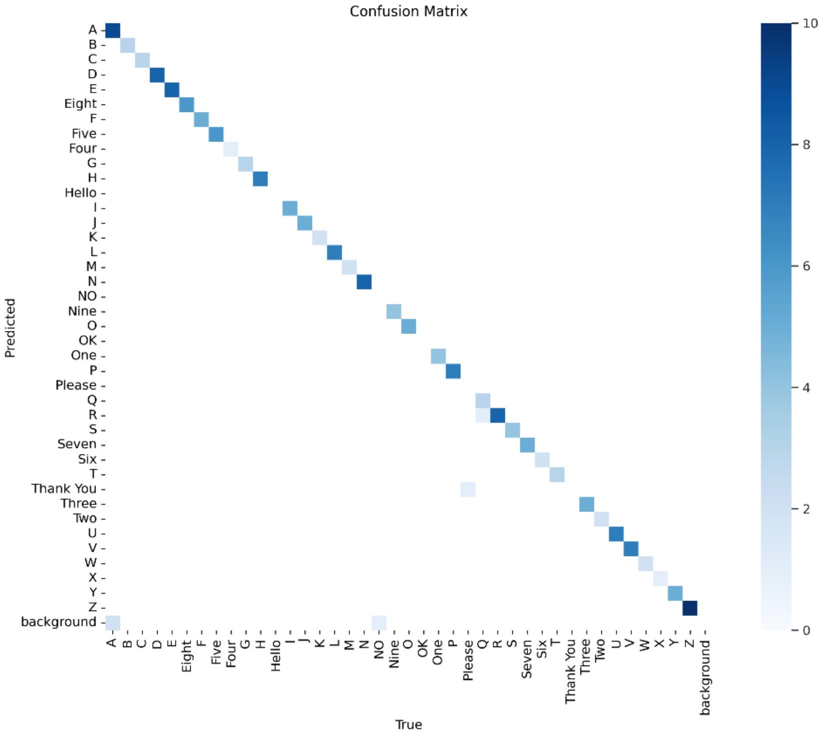

Figure 4. Confusion Matrix

Figure 4 presents the confusion matrix generated from the evaluation of the hand gesture classification model using data that includes letters (A–Z), numbers (0–9), and several words such as *"Hello"*, *"OK"*, *"Please"*, *"Thank You"*, along with *"background"* as an additional class.

The confusion matrix is a visual representation of classification performance, where the horizontal axis (True) indicates the actual labels, and the vertical axis (Predicted) represents the labels predicted by the model. Darker colors on the matrix indicate a higher number of predictions, with the color intensity scale shown on the right-hand side of the image. From the matrix, it can be observed that most predictions lie along the main diagonal, indicating that the model correctly classified the majority of the data. This suggests high accuracy across most classes, especially for letters like *"A"*, *"B"*, *"C"*, and words like *"Hello"* and *"Thank You"*, which appear as dark blue squares along the diagonal.

Nevertheless, a few misclassifications are visible as off-diagonal boxes, though they are relatively minimal. For example, there are classification errors between the letters *"N"* and *"M"*, and some confusion between the numbers *"One"* and *"Nine"*. These errors are likely due to the visual similarity of the gestures or noise in the input data.In addition, the *"background"* class was also predicted accurately in most cases, which is essential for minimizing false positives on non-gesture inputs.Overall, this confusion matrix demonstrates that the model achieves strong classification performance and is capable of distinguishing the majority of gesture classes with high accuracy. These results provide a solid foundation to claim the model's effectiveness in hand gesture recognition tasks.

3.2. Real-Time Implementation

To evaluate the model's performance in a real-world setting, a real-time implementation experiment was conducted using both a laptop camera and an external webcam. The objective of this experiment was to assess detection speed, gesture accuracy under real conditions, and the system's responsiveness when users performed SIBI alphabet gestures directly in front of the camera. Table 5 presents the hardware specifications used for testing the system in real-time conditions.

Table 5. Hardware Specification for Real-Time Testing

Component	Specification
Test Device	Laptop Lenovo Intel Core i7
GPU	Intel i7 Integrated Graphics
Camera	External Webcam
Video Input Resolution	640x640 pixels
Environment	Moderate Lighting Conditions

In this experiment, users were instructed to perform alphabet gestures (A to Z) and numerical gestures (0–9) sequentially in front of the camera. The system continuously captured images, detected hand objects, and classified the gestures into the corresponding SIBI letter categories using the trained YOLOv11 model. Detection results were visualized in a user interface displaying bounding boxes, gesture labels, and confidence scores in real-time.

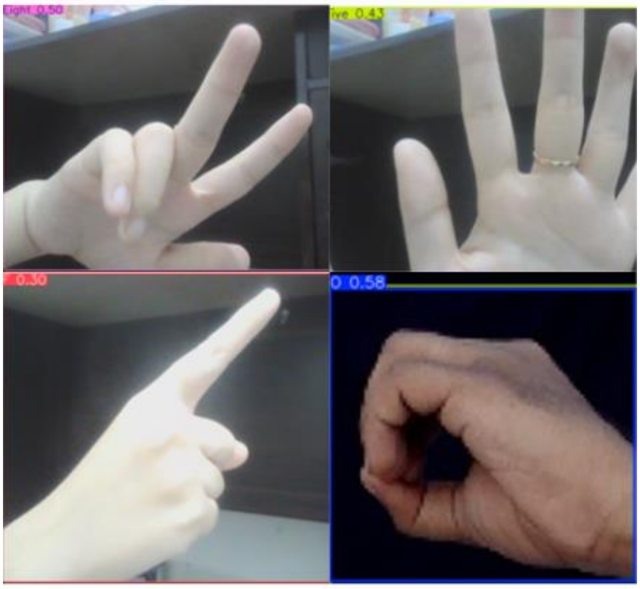

Figure 5. Real-Time SIBI Detection

Figure 5 displays four distinct hand gestures, each captured under different lighting conditions. In the top-left quadrant, the hand represents the gesture for the number *“eight”*, with the index and middle fingers extended. In the top-right quadrant, the gesture *“five”* is shown with all five fingers fully spread. The bottom-left quadrant presents the gesture *“one”*, where only the index finger is raised. Meanwhile, the bottom-right quadrant illustrates the gesture *“O”*, formed by connecting the thumb and index finger into a circle. Lighting conditions vary across the quadrants: the top images exhibit moderate lighting, the bottom-left image is taken under bright lighting, while the bottom-right image is captured in a low-light environment, resulting in less visible hand details.

Table 6. Experimental Result

Lighting Condition	Accuracy (%)	Average Detection Time (ms)	FPS
Bright Light	Laptop Lenovo Intel Core i7	32.1	31
Moderate Light	Intel i7 Integrated Graphics	33.5	30
Dim Light	External Webcam	35.7	28

As shown in Table 6, the model consistently maintained an accuracy above 90%, even under suboptimal lighting conditions and complex backgrounds. The average detection time ranged between 30 to 35 milliseconds, enabling the system to operate at real-time performance levels, achieving approximately 28–31 frames per second (FPS).

3.3. Comparative Discussion

Compared to earlier SIBI alphabet recognition systems using Convolutional Neural Networks (CNNs), the proposed YOLOv11-based model demonstrates significant improvements. For instance, a CNN-based SIBI detection system by (Sihananto, Safitri, Maulana, Fakhruddin, & Yudistira, 2023) reported accuracy of approximately 93.3%, yet depended on clean RGB images and lacked evaluation under varied lighting conditions. Similarly, (Alamsyah, 2024) achieved accuracy of 96.9% using a CNN-Nadam model, but this approach was limited to static image classification and did not support real-time inference or deployment on resource-constrained platforms.

In the context of earlier YOLO versions, YOLOv5-based systems have demonstrated strong performance in sign language recognition tasks. For example, Telugu sign language detection achieved mAP@0.5 of 98.1% with YOLOv5, though requiring extensive GPU resources and longer training times . However, YOLOv5 often struggles with subtle hand gestures and fine-grained spatial distinctions without architectural enhancements.

By contrast, the current YOLOv11 model achieves mAP@0.5 of 97%, with inference speeds up to 31 FPS on CPU-integrated GPU hardware, while maintaining high precision and recall across varying lighting conditions. These performance gains are attributable to YOLOv11's advanced architecture especially the C2PSA block, SPPF, and attention-enhanced feature fusion which strengthen spatial detail capture and recognition of small or occluded hand gestures (Khanam & Hussain, 2024).

Additionally, benchmarking studies confirm that YOLOv11 offers a balanced trade-off between model size, computational efficiency, and accuracy compared with prior YOLO versions. For instance, YOLOv11 achieves faster convergence and lower validation loss than YOLOv9/v10 on multiple datasets, with inference timings around 2.4 ms on GPU and strong performance on CPU scenarios (Jegham, Koh, Abdelatti, & Hendawi, 2025).

4. Conclusion

This study successfully developed a real-time SIBI alphabet gesture recognition system using the YOLOv11 object detection algorithm. The proposed model achieved high detection performance, with precision exceeding 95% and mAP@0.5 reaching approximately 97%. All loss metrics, including box loss, classification loss, and DFL loss, showed a consistent downward trend during training, indicating effective learning and minimal overfitting. The confusion matrix further demonstrated strong classification accuracy across most gesture classes, with very few misclassifications particularly between visually similar hand gestures. One of the key scientific contributions of this research is the implementation of YOLOv11 for SIBI alphabet recognition, which based on current literature represents one of the earliest applications of this state-of-the-art object detection model in the Indonesian context. Unlike previous studies that focused on international sign languages or utilized earlier YOLO versions and CNN-based methods, this research introduces an optimized, real-time model tailored to the structure and characteristics of SIBI. This provides a solid foundation for the development of localized, AI-based inclusive communication systems in Indonesia. In terms of practical contributions, the system demonstrates high accuracy, real-time performance on resource-limited hardware, and robustness across various lighting conditions making it suitable for integration into real-world assistive applications such as mobile apps, educational tools, or public service interfaces for the deaf and hard-of-hearing communities. For future work, it is recommended to expand the dataset with more diverse subjects and environmental variations, include dynamic sign sequences beyond the static alphabet, and explore deployment on mobile or embedded platforms such as Android or Raspberry Pi. The long-term goal is to build an end-to-end SIBI translation system that empowers inclusive communication and social participation for the deaf community in Indonesia.

REFERENCES

Aboud, H., Elsayaad, F., Gad, S. S., Abdelaziz, M., & Atia, A. (2024). Automated Rats Detection and Tracking for Behavioral Analysis in Biological Experiments. In *2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)* (pp. 307–312). IEEE. http://doi.org/10.1109/MIUCC62295.2024.10783596

Affairs, M. of S. (2023). *Data Penyandang Disabilitas 2022–2023*. Jakarta.

Agrawal, A. K., Kumar, J., Kumar, A., & Arvind, P. (2024). Analogous sign language communication using gesture detection. *Australian Journal of Electrical and Electronics Engineering*, *21*(4), 486–498.

http://doi.org/10.1080/1448837X.2024.2337495

Ahmed, A., Farhan, M., Eesaar, H., Chong, K. T., & Tayara, H. (2024). From Detection to Action: A Multimodal AI Framework for Traffic Incident Response. *Drones*, *8*(12), 741. http://doi.org/10.3390/drones8120741

Alamsyah, A. (2024). *Detection of Indonesian Sign Language System using Convolutional Neural Network (CNN) with Nadam Optimizer*. Atlantis Press International BV. http://doi.org/10.2991/978-94-6463-589-8

Ali, M. L., & Zhang, Z. (2024). The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. *Computers*, *13*(12), 336. http://doi.org/10.3390/computers13120336

Chen, G., Wang, F., Li, W., Hong, L., Conradt, J., Chen, J., … Knoll, A. (2022). NeuroIV: Neuromorphic Vision Meets Intelligent Vehicle Towards Safe Driving With a New Database and Baseline Evaluations. *IEEE Transactions on Intelligent Transportation Systems*, *23*(2), 1171–1183. http://doi.org/10.1109/TITS.2020.3022921

Das, A., Sayma, J., Barman, A. N., & Hasan, K. M. A. (2024). Application of YOLOv11 Classification for Efficient Waste Segmentation in Australia's Recycling Facilities. In *2024 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS)* (pp. 70–74). IEEE. http://doi.org/10.1109/AGERS65212.2024.10932955

Dignan, C., Perez, E., Ahmad, I., Huber, M., & Clark, A. (2022). An AI-based Approach for Improved Sign Language Recognition using Multiple Videos. *Multimedia Tools and Applications*, *81*(24), 34525–34546. http://doi.org/10.1007/s11042-021-11830-y

Gomez, A., & Arzuaga, E. (2024). Real Time American Sign Language Recognition Using Yolov6 Model (pp. 343–353). http://doi.org/10.1007/978-3-031-67447-1_25

Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., & Xu, C. (2022). CMT: Convolutional Neural Networks Meet Vision Transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12165–12175). IEEE. http://doi.org/10.1109/CVPR52688.2022.01186

He, L., Zhou, Y., Liu, L., & Ma, J. (2024). Research and Application of YOLOv11-Based Object Segmentation in Intelligent Recognition at Construction Sites. *Buildings*, *14*(12), 3777. http://doi.org/10.3390/buildings14123777

Jegham, N., Koh, C. Y., Abdelatti, M., & Hendawi, A. (2025). YOLO Evolution: A Comprehensive Benchmark and Architectural Review of YOLOv12, YOLO11, and Their Previous Versions. Retrieved from http://arxiv.org/abs/2411.00201

Khanam, R., & Hussain, M. (2024). YOLOv11: An Overview of the Key Architectural Enhancements. Retrieved from http://arxiv.org/abs/2410.17725

Liu, B., & Li, X. (2024). An Improved YOLOv11 Model for Detecting the Metal Roofing Tiles alongside the Railways. In *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)* (pp. 195–199). IEEE. http://doi.org/10.1109/ICAIRC64177.2024.10900077

Liu, Y., Nand, P., Hossain, M. A., Nguyen, M., & Yan, W. Q. (2023). Sign language recognition from digital videos using feature pyramid network with detection transformer. *Multimedia Tools and Applications*, *82*(14), 21673–21685. http://doi.org/10.1007/s11042-023-14646-0

Ong, S. C. W., & Ranganath, S. (2005). Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(6), 873–891. http://doi.org/10.1109/TPAMI.2005.112

Parico, A. I. B., & Ahamed, T. (2021). Real Time Pear Fruit Detection and Counting Using YOLOv4 Models and Deep SORT. *Sensors*, *21*(14), 4803. http://doi.org/10.3390/s21144803

Rodríguez-Lira, D.-C., Córdova-Esparza, D.-M., Álvarez-Alvarado, J. M., Romero-González, J.-A., Terven, J., & Rodríguez-Reséndiz, J. (2024). Comparative Analysis of YOLO Models for Bean Leaf Disease Detection in Natural Environments. *AgriEngineering*, *6*(4), 4585–4603. http://doi.org/10.3390/agriengineering6040262

Roy, A. M., Bhaduri, J., Kumar, T., & Raj, K. (2023). WilDect-YOLO: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Ecological Informatics*, *75*, 101919. http://doi.org/10.1016/j.ecoinf.2022.101919

Safitri, M., Yuniarno, E. M., & Rachmadi, R. F. (2024). Indonesian Sign Language (SIBI) Recognition and Extraction Using Convolutional Neural Networks - Symmetric Deletion Spelling Correction. In *2024 International Seminar on Intelligent Technology and Its Applications (ISITIA)* (pp. 220–225). IEEE. http://doi.org/10.1109/ISITIA63062.2024.10667714

Sani, A. R., Zolfagharian, A., & Kouzani, A. Z. (2024). Automated defects detection in extrusion 3D printing using YOLO models. *Journal of Intelligent Manufacturing*. http://doi.org/10.1007/s10845-024-02543-8

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. http://doi.org/10.1007/s11263-019-01228-7

Shivaprasad Yadav, S. G., Itagi, S., Krishna Suresh, B. V. N. V, K.L, H., & A C, R. (2023). Human Illegal Activity Recognition Based on Deep Learning Techniques. In *2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS)* (pp. 01–07). IEEE. http://doi.org/10.1109/ICICACS57338.2023.10099857

Sihananto, A. N., Safitri, E. M., Maulana, Y., Fakhruddin, F., & Yudistira, M. E. (2023). Indonesian Sign Language Image Detection Using Convolutional Neural Network (CNN) Method. *Inspiration: Jurnal Teknologi Informasi Dan Komunikasi*, *13*(1), 13–21. http://doi.org/10.35585/inspir.v13i1.37

Statistik, B. P. (2022). *Susenas 2022: Penyandang Disabilitas Menurut Jenis Disabilitas dan Provinsi*. Jakarta.

Sümbül, H. (2024). A Novel Mems and Flex Sensor-Based Hand Gesture Recognition and Regenerating System Using Deep Learning Model. *IEEE Access*, *12*, 133685–133693. http://doi.org/10.1109/ACCESS.2024.3448232

Tourani, A., Soroori, S., Shahbahrami, A., & Akoushideh, A. (2021). Iranis: A Large-scale Dataset of Iranian Vehicles License Plate Characters. In *2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA)* (pp. 1–5). IEEE. http://doi.org/10.1109/IPRIA53572.2021.9483461

Xiao, B., & Kang, S.-C. (2021). Development of an Image Data Set of Construction Machines for Deep Learning Object Detection. *Journal of Computing in Civil Engineering*, *35*(2). http://doi.org/10.1061/(ASCE)CP.1943-5487.0000945

Yang, J., Tian, T., Liu, Y., Li, C., Wu, D., Wang, L., & Wang, X. (2024). A Rainy Day Object Detection Method Based on YOLOv11 Combined with FFT and MF Model Fusion. In *2024 International Conference on Advanced Control Systems and Automation Technologies (ACSAT)* (pp. 246–250). IEEE. http://doi.org/10.1109/ACSAT63853.2024.10823725

Zanevych, Y., Yovbak, V., Basystiuk, O., Shakhovska, N., Fedushko, S., & Argyroudis, S. (2024). Evaluation of Pothole Detection Performance Using Deep Learning Models Under Low-Light Conditions. *Sustainability*, *16*(24), 10964.

http://doi.org/10.3390/su162410964

Zhang, C., Cheng, H., Wu, R., Ren, B., Zhu, Y., & Peng, N. (2024). Development of a Traffic Congestion Prediction and Emergency Lane Development Strategy Based on Object Detection Algorithms. *Sustainability*, *16*(23), 10232. http://doi.org/10.3390/su162310232