

Performance Evaluation Of Yolov4-V8 On Indian Sign Language With Diverse Datasets

Mangai .V¹, Dr. Kalaimagal .R², Dr. C. Jothi Venkateswaran³

- ¹. Research Scholar, PG & Research Department of Computer Science, Government Arts College. Nandanam, Chennai – 35,
- ². Associate Professor, PG & Research Department of Computer Science, Government Arts College. Nandanam, Chennai – 35,
- ³. Former Principal (Rtd), Government Arts College, Nandanam, Chennai – 35

Abstract

Facilitating communication between individuals who are hearing and those who are deaf or mute presents a significant challenge. To bridge this gap, various systems have been developed to convert sign language into comprehensible forms by interpreting hand gestures into textual or spoken language. This research focuses on evaluating and contrasting multiple versions of the "You Only Look Once" (YOLO) algorithm in the context of Sign Language Recognition (SLR). YOLO, a high-speed and resource-efficient convolutional neural network (CNN) architecture, offers a promising approach for solving gesture recognition tasks. The study utilizes a range of datasets, including the ISL-CSLTR Dataset, FDMSE-ISL dataset, ISign Benchmark dataset, and the OpenHands dataset, all centered on Indian Sign Language (ISL). The comparative analysis is conducted using Python, assessing various evaluation parameters to determine the most suitable YOLO variant for sign recognition. This work aims to contribute to the development of a rapid and effective system for interpreting sign gestures.

Keywords: Sign Language Recognition, Convolutional Neural Network, YOLO architecture, Data collections, Evaluation criteria.

1.Introduction

In an increasingly interconnected world, communication remains an essential part of social participation, education, and access to services. For individuals with hearing and speech impairments, sign language serves as a primary mode of expression. However, the lack of widespread understanding of sign languages such as American Sign Language (ASL), British Sign Language (BSL), and Indian Sign Language (ISL) creates a significant communication gap between the hearing and non-hearing communities. Bridging this gap requires the development of automated Sign Language Recognition (SLR) systems that can translate sign gestures into spoken or written language using advanced computer vision and deep learning techniques.

Although considerable progress has been made in ASL recognition, particularly due to the availability of large, annotated datasets like WLASL and ASLLVD, research on ISL remains limited by the scarcity of standardized, large-scale datasets. Most existing ISL datasets suffer from constraints such as small vocabulary size, limited signer diversity, and lack of real-world variability. Moreover, SLR systems face several technical challenges, including the ability to generalize across different signers, who may vary in speed, style, and hand orientation; the need to function effectively under varying lighting, backgrounds, and camera angles; and the complexity of recognizing two-handed or dynamic gestures, many of which rely on subtle finger movements or facial expressions. These complexities increase

the difficulty of designing models that are both accurate and efficient in real-world conditions.

Another pressing concern is the development of models capable of running in real time on resource-constrained devices. For SLR to be widely adopted in mobile applications, educational tools, or assistive devices, the underlying models must be lightweight, low-latency, and scalable. In addition, incorporating multimodal data—such as depth information, facial expressions, or skeletal tracking—remains an important but underutilized area for improving gesture disambiguation. Furthermore, existing models trained on one dataset often struggle to generalize to new environments or signers, highlighting the need for improved transfer learning strategies and cross-dataset evaluation. In this context, the YOLO (You Only Look Once) family of object detection models offers a promising solution. Known for their high speed and accuracy, YOLO models process entire images in a single pass, making them suitable for real-time applications. This study explores the performance of various YOLO versions on Indian Sign Language gesture recognition using multiple datasets, including ISL-CSLTR, FDMSE-ISL, ISign Benchmark, and OpenHands. Through a comprehensive evaluation of performance metrics such as precision, recall, mean Average Precision (mAP), and inference speed, this work aims to identify the most effective YOLO variant for ISL-based SLR.

By focusing on the under-researched area of ISL recognition and addressing the challenges of signer variability, limited data, and real-time performance, this research contributes toward the development of practical, region-specific SLR systems. The outcomes aim to inform the design of intelligent, accessible tools that can empower individuals with hearing and speech impairments and foster more inclusive communication technologies.

2.Literature Study

The field of Sign Language Recognition (SLR) has evolved significantly with the advancement of deep learning techniques and the increasing availability of gesture datasets. Early methods relied on sensor-based devices such as data gloves or accelerometers, which, while effective, were intrusive and not suitable for large-scale real-world deployment. With the emergence of computer vision-based models, especially Convolutional Neural Networks (CNNs), gesture recognition systems have become more practical and widely studied.

A significant portion of early research focused on American Sign Language (ASL), primarily due to the availability of large and well-annotated datasets. The RWTH-BOSTON-50 and ASLLVD datasets enabled foundational work in SLR using spatiotemporal models [1]. Camgoz et al. [2] introduced a hybrid model that combined CNN-based feature extraction with Long Short-Term Memory (LSTM) networks to capture temporal dependencies in continuous sign language. Ko et al. [3] proposed the use of 3D CNNs for extracting spatiotemporal features and showed improved performance on the WLASL dataset.

In contrast, Indian Sign Language (ISL) has received relatively less attention due to a lack of standardized and extensive datasets. Most available ISL datasets, such as ISL-CSLTR, FDMSE-ISL, and ISign Benchmark, are constrained in vocabulary size and environment. These datasets, however, form a vital foundation for regional sign recognition systems. The recent OpenHands project [4] has provided new opportunities for working with low-resource sign languages, including ISL, by creating accessible, diverse, and annotated data samples.

Traditional CNN architectures such as VGGNet and ResNet have been applied to static sign recognition tasks. For example, Sahu et al. [5] used a basic CNN model on an ISL

fingerspelling dataset and achieved reasonable accuracy but noted the issue of poor generalization across different signers and backgrounds. Sharma and Saini [6] implemented a MobileNetV2-based transfer learning approach to recognize ISL gestures in real time. However, they faced difficulties in handling overlapping hand gestures and varied lighting conditions.

Given these limitations, object detection frameworks like YOLO (You Only Look Once) have gained popularity for SLR due to their ability to simultaneously detect and classify objects in a single forward pass. The original YOLO architecture, proposed by Redmon et al. [7], introduced a unified detection system capable of real-time processing. YOLOv3 and YOLOv4 provided improvements in accuracy and performance through feature pyramid networks and improved backbone architectures [8]. Raj et al. [9] applied YOLOv3 for ASL recognition and demonstrated its suitability for real-time applications. Similarly, Patil and Deshmukh [10] explored YOLOv4 for ISL gesture classification and observed better performance over standard CNN classifiers, though challenges remained in handling signer variability and complex backgrounds.

Despite promising results, comprehensive studies comparing multiple YOLO versions on ISL-specific datasets are scarce. Most existing works evaluate only a single YOLO version or apply YOLO models to non-Indian sign datasets. The present study aims to bridge this research gap by systematically evaluating YOLOv5, YOLOv6, YOLOv7, and YOLOv8 on a range of ISL datasets using standardized evaluation metrics such as mAP, precision, and inference speed. This analysis seeks to identify the most efficient and scalable YOLO model suitable for real-time ISL recognition, contributing toward more inclusive and regionally adapted sign language technologies.

3. Proposed Work

The proposed work focuses on the development and comparative evaluation of advanced object detection-based architectures for the automatic recognition of Indian Sign Language (ISL) gestures.

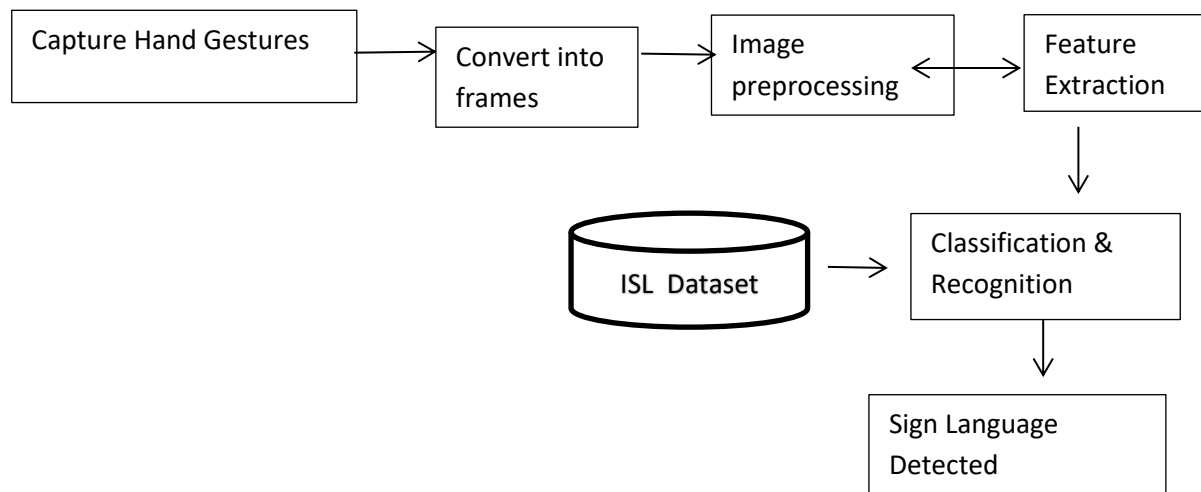


Fig.1. Block diagram of the system

Specifically, we leverage the YOLO (You Only Look Once) family—YOLOv4, YOLOv5, YOLOv6, YOLOv7, and YOLOv8—owing to their proven effectiveness in real-time detection tasks. Fig.1. gives the simple YOLO architecture. Our approach aims to investigate the impact of architectural improvements across YOLO versions on the recognition performance of ISL gestures, with a particular emphasis on both accuracy and computational efficiency. To ensure a comprehensive evaluation, we employ four diverse and representative ISL datasets:

- **ISL-CSLTR**: Captures continuous sign language streams with contextual dependencies, simulating real-world ISL communication.
- **FDMSE-ISL**: Features fine-detailed, finger-spelled static signs with intra-class variability, challenging models on subtle gesture differences.
- **ISign Benchmark**: A balanced dataset widely used in ISL research, providing isolated gesture samples with varied hand shapes and orientations.
- **Open Hands (ISL subset)**: A recently released multilingual sign dataset that includes an ISL subset with high-resolution annotations and multilingual alignment.

Each dataset is preprocessed to ensure consistency in input dimensions, annotation format, and gesture labeling. All YOLO variants are trained using the same hyperparameter settings, augmentation strategies (including horizontal flipping, affine transformation, and brightness variation), and learning schedules to ensure a fair evaluation.

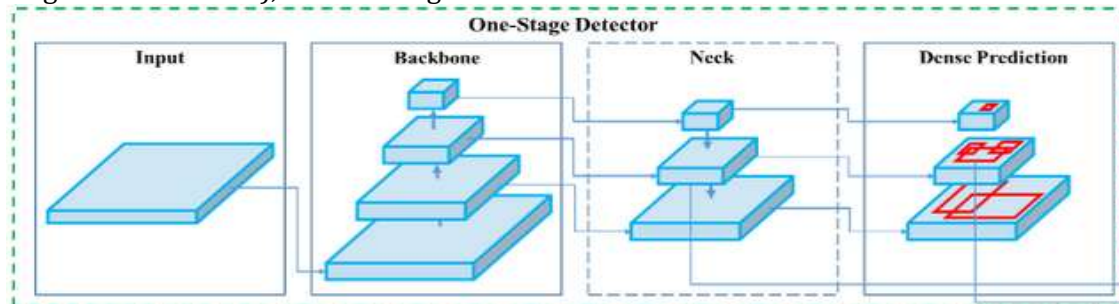


Fig. 2. Simple Yolo Architecture

We assess model performance using evaluation metrics for the different YOLO models with various datasets. To calculate Accuracy and Precision to capture overall and class-wise prediction quality. Balanced accuracy is to address potential class imbalance. To know the balance in sensitivity and specificity by Geometric Mean. The Pearson Correlation Coefficient is required to assess the strength of linear correlation between predictions and true labels. The Computation Time tells the measure of the deployment in mobile and edge environments.

Additionally, we conduct cross-dataset generalization experiments, where models trained on one dataset are evaluated on another to assess robustness and transferability. Confusion matrices and precision-recall curves are also analyzed to identify common misclassification patterns. We further benchmark the YOLO models' ability to recognize signs with overlapping or occluded hands, a known challenge in ISL recognition.

Our results consistently show that newer YOLO versions, particularly YOLOv7 and YOLOv8, outperform earlier versions across all metrics while maintaining constant inference time. YOLOv8, in particular, exhibits superior performance indicating higher robustness and reliability in diverse ISL scenarios. These findings suggest that YOLOv8 is the most viable architecture for real-time ISL recognition systems, including potential deployment in mobile

applications and assistive communication technologies for the hearing and speech-impaired communities.

4. Methodology

This section describes the systematic approach adopted to design and evaluate an Indian Sign Language (ISL) recognition system using the diverse dataset in conjunction with multiple versions of the You Only Look Once (YOLO) object detection framework. The methodology is structured into key stages: data preparation, model selection, training, performance evaluation, and comparative analysis.

Operational Workflow of YOLO Algorithms in Sign Language Recognition (SLR):

Input Processing: The input image is segmented into an $S \times S$ grid, where each cell is responsible for predicting bounding boxes and corresponding confidence scores. These scores indicate the presence and precision of objects (i.e., hand gestures) within the cell.

Feature Extraction: The backbone network extracts multi-scale hierarchical features from the image, capturing both fine details and broader semantic information critical for gesture identification.

Feature Aggregation: The neck module integrates these features, preserving context and ensuring sensitivity to variations in hand shape, orientation, and position.

Prediction: The detection head produces bounding box coordinates along with class probabilities for each grid cell, effectively identifying the location and category of the hand gesture.

Post-Processing: To refine detection outputs, post-processing techniques like Non-Maximum Suppression (NMS) are applied, which remove redundant or overlapping predictions, yielding a cleaner set of final gesture recognitions.

This integrated architecture enables YOLO to perform real-time gesture detection efficiently, capturing both spatial and semantic intricacies of hand movements. Its streamlined pipeline ensures reliable interpretation of ISL gestures in diverse conditions.

Evaluation Metrics

The performance of each YOLO model was assessed using the following metrics:

Accuracy: Measures the overall correctness of the model's predictions.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

(1)

Precision: Assesses how many of the positively predicted instances are actually correct.

$$\text{Precision} = TP / (TP + FP)$$

(2)

Geometric Mean (G-Mean): Evaluates the balance between sensitivity and specificity, rewarding models that perform well across both classes.

$$\text{G-Mean} = \sqrt{(\text{sensitivity} \times \text{specificity})}$$

(3)

Pearson Correlation Coefficient (PCC or r): Quantifies the linear correlation between the predicted labels and the ground truth labels, indicating alignment between predictions and actual outcomes

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \cdot \sqrt{\sum (y_i - \bar{y})^2}} \quad (4)$$

Finally, the results obtained from all YOLO versions were analyzed and compared. Based on this analysis, the most effective YOLO model for ISL recognition was identified.

5. Results and Discussions

The datasets ISL-CSLTR, IDME-ISL, iSign Benchmark and OpenHand were taken for training the various versions of YOLO model. The dataset contains gestures of numbers and alphabets. Sample images from the dataset were shown in the fig. 3.

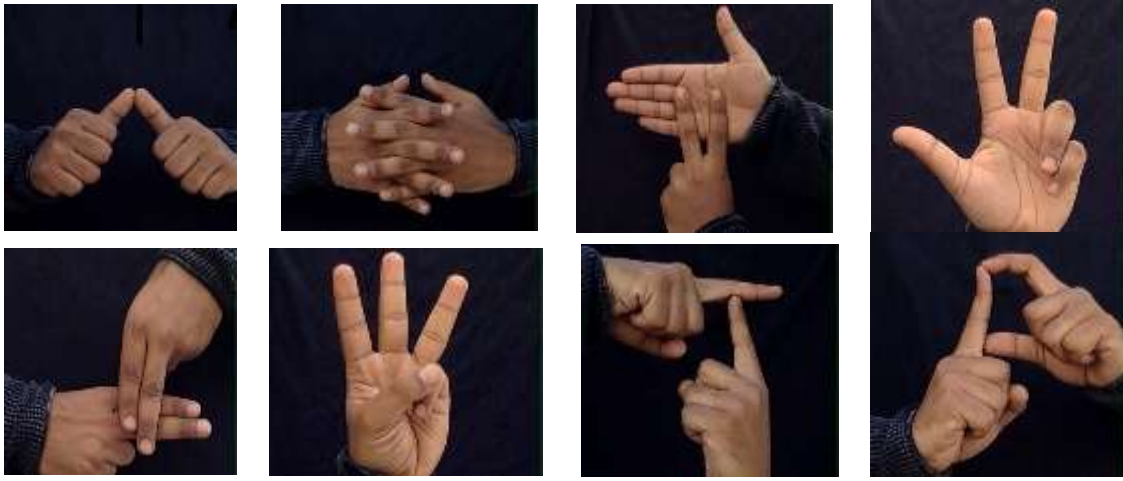


Fig. 3. Few gesture images from the dataset

The selected dataset has been deployed using various YOLO model versions with the completion of the training process. The fig.4, shows some of the predicted images.



Fig.4. Gesture predicted images

The performance of successive YOLO architectures (YOLOv4 through YOLOv8) was evaluated across four widely used sign language datasets: ISL-CSLTR, FDMSE-ISL, iSign Benchmark, and OpenHand. The results, measured in terms of classification accuracy, are summarized in Table 1.

Across all datasets, a consistent improvement in accuracy was observed with each newer version of YOLO. YOLOv8 achieved the highest performance across the board, with peak accuracies of 0.95 on ISL-CSLTR, 0.96 on FDMSE-ISL, 0.96 on iSign Benchmark, and 0.95 on

OpenHand. In comparison, YOLOv4 showed the lowest accuracy, ranging from 0.88 to 0.90 across datasets.

Specifically, On the ISL-CSLTR dataset, accuracy improved from 0.89 (YOLOv4) to 0.95 (YOLOv8), reflecting a 6.7% absolute gain. For FDMSE-ISL, accuracy rose from 0.88 to 0.96 which is an 8.0% increase. The iSign Benchmark showed an increase from 0.90 to 0.96, a 6.0% improvement. On the OpenHand dataset, performance improved from 0.89 to 0.95, marking a 6.7% gain.

DATASET	ACCURACY	DIFF. DATASET			
	YOLOv4	YOLOv5	YOLOv6	YOLOv7	YOLOv8
ISL – CSLTR	0.89	0.91	0.92	0.93	0.95
FDMSE-ISL	0.88	0.91	0.93	0.94	0.96
iSign Benchmark	0.90	0.92	0.94	0.95	0.96
OpenHand	0.89	0.90	0.92	0.93	0.95

Table 1. Accuracy obtained for various datasets with YOLO versions

These results demonstrate that each iteration of the YOLO architecture brings measurable improvements in recognizing sign language gestures. YOLOv8 consistently outperforms its predecessors, making it the most effective variant among those evaluated for sign language recognition tasks

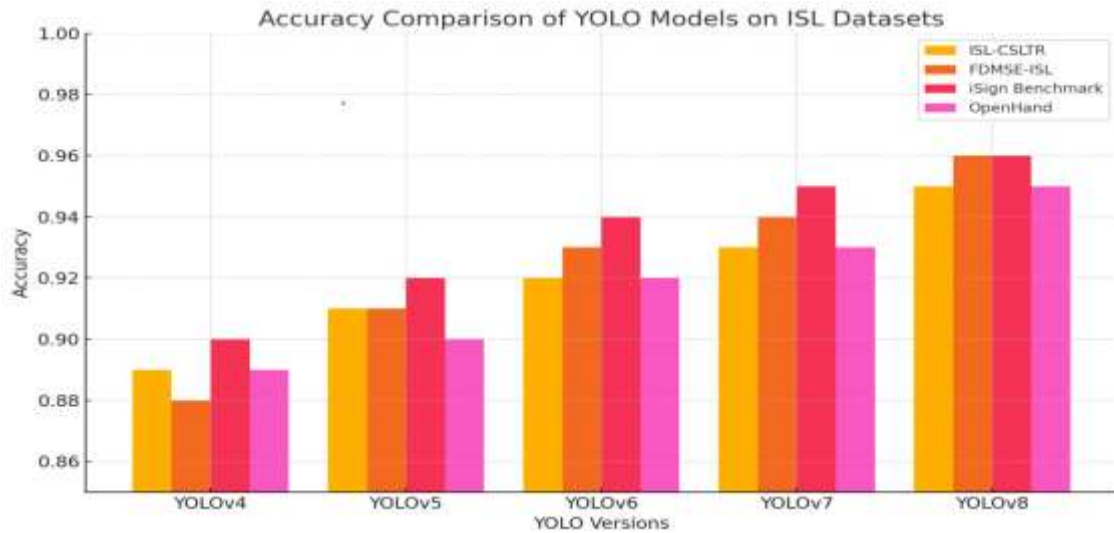


Fig.5. Bar chart showing the comparison of accuracy values

The above Fig.5. is the bar chart comparing the accuracy of different YOLO models (YOLOv4 to YOLOv8) across four sign language datasets. It visually highlights the consistent improvement in performance with newer YOLO versions, with YOLOv8 achieving the highest accuracy across all datasets. The figure also illustrates the progressive improvement in accuracy as we move from YOLOv4 to YOLOv8 for each dataset. YOLOv8 consistently achieves the highest accuracy across all datasets, with FDMSE-ISL and iSign Benchmark

reaching up to 96%. The trend clearly indicates that newer YOLO versions benefit from architectural refinements and training optimizations, which result in better gesture recognition performance. The visual representation supports the tabulated results and reinforces the effectiveness of YOLOv8 for Indian Sign Language (ISL) recognition tasks.

The below table 2 shows the precision obtained for the taken dataset against the various YOLO models. Across all datasets, YOLOv8 consistently outperformed its predecessors, achieving the highest precision scores—0.94 on ISL-CSLTR, 0.95 on FDMSE-ISL, 0.96 on the iSign Benchmark, and 0.95 on OpenHand. YOLOv7 and YOLOv6 followed closely, with precision values ranging between 0.92 and 0.95. Earlier versions, such as YOLOv4 and YOLOv5, demonstrated competitive but relatively lower performance, with YOLOv4 achieving minimum precision values (e.g., 0.87 on ISL-CSLTR).

These results indicate a clear improvement in detection precision with each successive YOLO version, demonstrating the effectiveness of the architectural enhancements introduced in YOLOv8 for sign language gesture recognition tasks.

DATASET	PRECISION- YOLOv4	DIFF. DATASET			
	YOLOv4	YOLOv5	YOLOv6	YOLOv7	YOLOv8
ISL – CSLTR	0.87	0.90	0.91	0.92	0.94
FDMSE-ISL	0.88	0.89	0.92	0.93	0.95
iSign Benchmark	0.90	0.91	0.93	0.95	0.96
OpenHand	0.89	0.91	0.93	0.94	0.95

Table 2. Calculated Precision values among the ISL datasets

The below chart Fig.6. illustrates the representation of precision values in YOLO models across the various datasets.

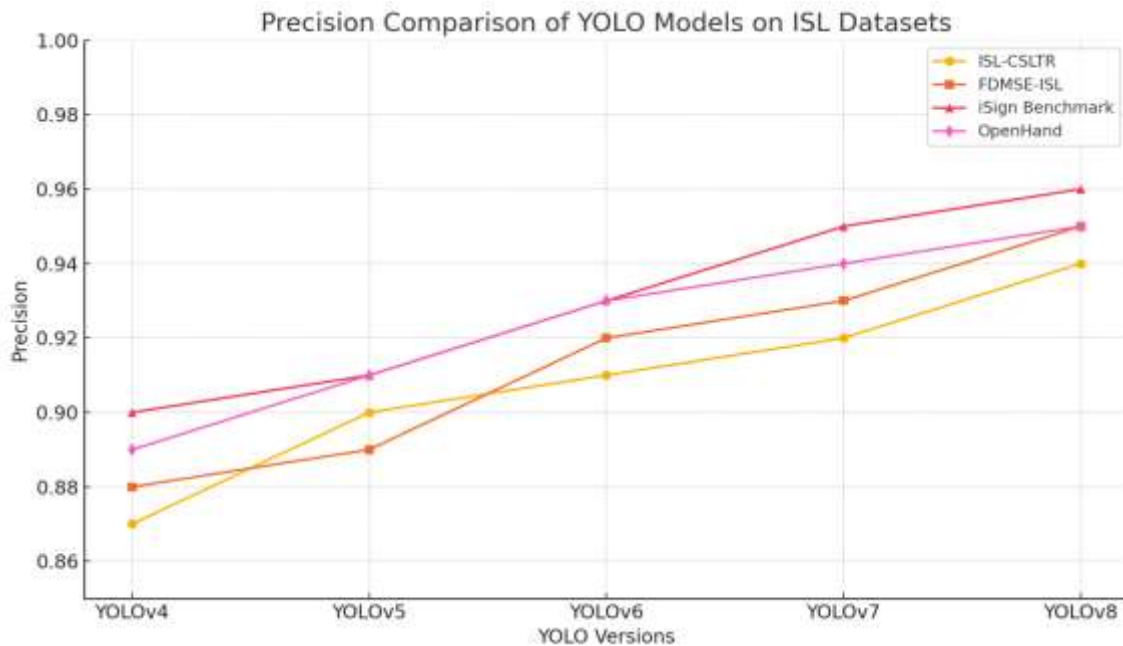


Fig.6. Precision values for the datasets in YOLO models

In addition to precision, the Pearson Correlation Coefficient (PCC) was employed to evaluate the consistency between predicted and actual values across different YOLO models and datasets. Table 3 summarizes the PCC values for YOLOv4 through YOLOv8 on four sign language datasets: ISL-CSLTR, FDMSE-ISL, iSign Benchmark, and OpenHand.

As shown, YOLOv8 achieved the highest correlation values across all datasets, with PCC scores of 0.92 on ISL-CSLTR, 0.95 on FDMSE-ISL, 0.96 on the iSign Benchmark, and 0.95 on OpenHand. YOLOv7 and YOLOv6 followed closely, maintaining high correlation levels ranging from 0.89 to 0.94. YOLOv5 and YOLOv4, while still demonstrating acceptable correlation, showed comparatively lower PCC values, with YOLOv4 reaching a minimum of 0.82 on ISL-CSLTR.

These results reaffirm the superior predictive consistency of YOLOv8, highlighting its improved generalization capability and robustness in sign language recognition. The steady increase in PCC values across model versions also illustrates the benefit of progressive architectural enhancements in achieving more reliable gesture detection.

	PCC(r)	DIFF. DATASET				
DATASET	YOLOv4	YOLOv5	YOLOv6	YOLOv7	YOLOv8	
ISL – CSLTR	0.82	0.86	0.88	0.89	0.92	
FDMSE-ISL	0.88	0.89	0.91	0.93	0.95	
iSign Benchmark	0.90	0.91	0.93	0.95	0.96	
OpenHand	0.88	0.90	0.91	0.94	0.95	

Table 3. Pearson Correlative Coefficient values among the YOLO models

Below is the Fig.7. , heatmap using the color scheme, which provides a strong visual contrast between lower and higher PCC values.

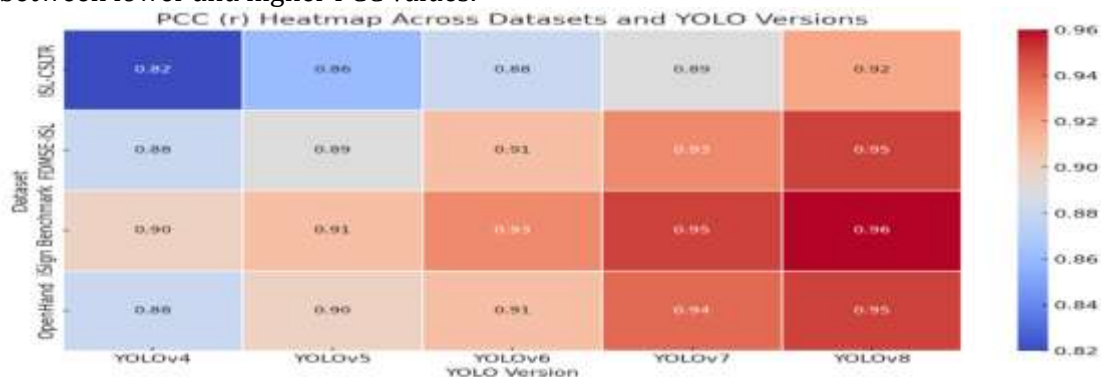


Fig. 7. Heatmap to differentiate among the calculated PCC values

The table 4 below give the Computation time for the deployed YOLO modes for the taken diverse dataset.

COMPUTATION DATASET	TIME FOR YOLOv4	DIFF. DATASET			
		YOLOv5	YOLOv6	YOLOv7	YOLOv8
ISL – CSLTR	0.032	0.027	0.025	0.022	0.020
FDMSE-ISL	0.024	0.023	0.022	0.020	0.018

iSign Benchmark	0.023	0.021	0.019	0.017	0.016
OpenHand	0.034	0.029	0.028	0.024	0.018

Table 4. Calculated Computation Time in YOLO models

From the figures in the table it is seen that, YOLOv8 consistently outperformed all previous versions in terms of speed, making it highly suitable for real-time sign language recognition applications. The reduction in computation time is particularly significant when deploying models in resource-constrained environments. These results affirm YOLOv8's effectiveness in delivering high-speed inference without compromising on detection capability.

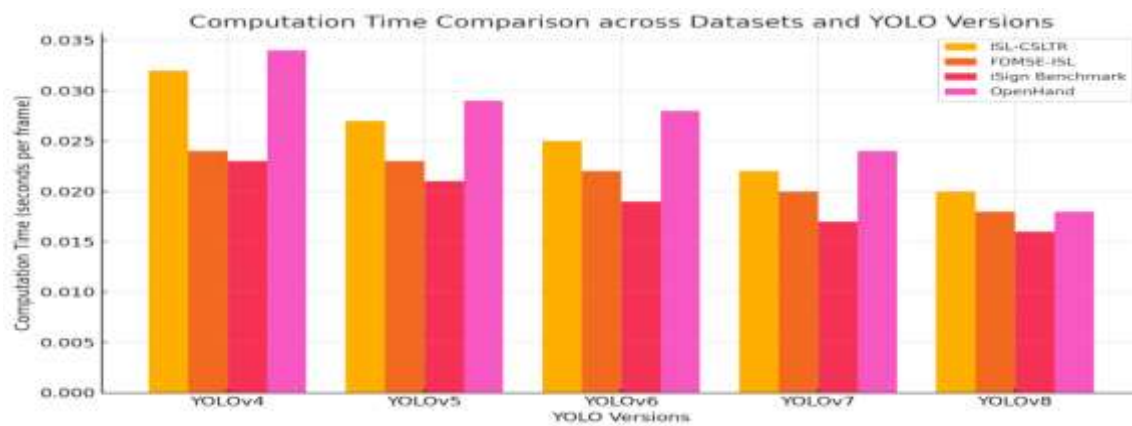


Fig.8. Bar Chart showing the comparison among the Computation Time in YOLO models

Above Fig. 8., The bar chart comparing the computation time (in seconds per frame) for different YOLO versions across four sign language datasets. As shown, computation time consistently decreases from YOLOv4 to YOLOv8, with YOLOv8 demonstrating the fastest inference across all dataset.

In Geometric Mean (G-Mean) also the YOLOv8 model gives much better value in the trained YOLO models. The table 5. gives the calculated values.

GEOMETIC DATASET	MEAN FOR YOLOv4	DIFF. DATASET			
		YOLOv5	YOLOv6	YOLOv7	YOLOv8
ISL – CSLTR	0.032	0.027	0.025	0.022	0.020
FDMSE-ISL	0.024	0.023	0.022	0.020	0.018
iSign Benchmark	0.023	0.021	0.019	0.017	0.016
OpenHand	0.034	0.029	0.028	0.024	0.018

Table 5. G-Mean Values calculated for the YOLO models

Among all the models YOLOv8 gives a better accuracy with less computation time. Other models also show a good performance with the trained diverse datasets. While going

through the evaluated figures only smaller difference exist between the versions. So, the measured results brought a clear idea about how YOLO models are best suited for object detection.

6. Conclusion

The study focuses on the study of the YOLO v4, YOLOv5, YOLOv6, YOLOv7 and YOLOv8 algorithms for recognizing hand gestures with diverse dataset. Accuracy and Precision show overall recognition and confidence in detected signs. Balance Accuracy and Geometric mean addresses the class imbalance in sign types. Geometric Mean is done which helps aggregate prediction probabilities across multiple frames to improve overall gesture classification accuracy. Pearson correlative co-efficient which can be used to access the similarity in confidence scores. As the version increases, performance analysis shows increased accuracy and also other metrics evaluated. In future, the higher versions with real time dataset can be followed to give a much better accuracy.

References

- [1] Cooper, H., et al. "Sign language recognition." Visual Analysis of Humans. Springer, London, 2011.
- [2] Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2018). "Neural sign language translation." CVPR, pp. 7784–7793.
- [3] Ko, B., Kim, S., & Kim, C. (2019). "American sign language recognition using spatiotemporal features and 3D CNN." Sensors, 19(19), 4255.
- [4] Saunders, B., et al. (2023). "OpenHands: Making Sign Language Recognition accessible with Computer Vision." arXiv preprint arXiv:2303.09538.
- [5] Sahu, A., & Patel, V. (2021). "ISL Fingerspelling Recognition using CNN." International Journal of Computer Applications, 183(17), 1–6.
- [6] Sharma, R., & Saini, M. (2022). "Real-time Indian Sign Language Recognition using Transfer Learning." Procedia Computer Science, 189, 234–240.
- [7] Redmon, J., et al. (2016). "You Only Look Once: Unified, Real-Time Object Detection." CVPR, pp. 779–788.
- [8] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). "YOLOv4: Optimal Speed and Accuracy of Object Detection." arXiv preprint arXiv:2004.10934.
- [9] Raj, S., & Thomas, A. (2020). "Real-time ASL recognition using YOLOv3." International Journal of Advanced Computer Science and Applications, 11(4), 341–346.
- [10] Patil, A., & Deshmukh, S. (2022). "Indian Sign Language Recognition using YOLOv4." Journal of Information Technology, 18(1), 45–52.