SIGN LANGUAGE DETECTION USING YOLOv10

Lalitha Madanbhavi
School of Computer Science
KLE Technological University
Hubli, Karnataka
lalita@kletech.ac.in

Devaraj Hireraddi
School of Computer Science
KLE Technological University
Hubli, Karnataka
devarajhireraddi777@gmail.com

Nitin Nagaral
School of Computer Science
KLE Technological University
Hubli, Karnataka
nitin.nagaral@gmail.com

Kushalagouda Patil
School of Computer Science
KLE Technological University
Hubli, Karnataka
kushalpatil484@gmail.com

Girish Dongrekar
School of Computer Science
KLE Technological University
Hubli, Karnataka
girishdongrekar123@gmail.com

Mahendra Iraddi
School of Computer Science
KLE Technological University
Hubli, Karnataka
mahendrairaddi07@gmail.com

Arundati Aralimatti
School of Computer Science
KLE Technological University
Hubli, Karnataka
arundati.aralimatti@kletech.ac.in

Chetan Patil
School of Computer Science
R. V. College of Engineering
Bengaluru, Karnataka
Chetanmpatil.cs23@rvce.edu.in

***Abstract*—Communication barriers pose significant challenges for individuals who are deaf, impacting their ability to interact seamlessly in society. American Sign Language (ASL) serves as a vital communication medium within this community, yet it remains largely unfamiliar to the general populace. This gap hampers access to essential services and social interactions. Leveraging advancements in deep learning and computer vision technologies, this project introduces a real-time ASL letter detection system utilizing the YOLOv10 model, renowned for its efficiency and accuracy in object detection. The system captures live video feed via a webcam, detects ASL letters, and provides visual feedback through annotated video frames. Our model, trained on a custom dataset encompassing diverse hand positions, orientations, and lighting conditions, demonstrates high accuracy and robustness. The methodology includes data collection and preprocessing, model architecture and training, and real-time detection implementation. Evaluation metrics such as precision, recall, mean Average Precision (mAP), and accuracy affirm the model's effectiveness, with an accuracy of 92.86%. This system offers significant potential for applications in education, healthcare, and public services, promoting inclusivity and accessibility. Future work aims to enhance the model's performance by expanding the ASL dataset's size and diversity, thereby improving its ability to generalize across various real-world scenarios.**

***Index Terms*—ASL, YOLOv10, realtime detection, OpenCV, PyTorch**

I. Introduction

Communication barriers pose significant challenges for individuals who are deaf, affecting their ability to interact seamlessly with others in society. ASL is a predominant means of communication among these individuals, providing a visual language that relies on hand gestures, facial expressions, and body language. Despite its widespread use within the deaf community, ASL is not universally understood by the general population, leading to a significant communication gap. This gap can impede access to essential services, social interactions, and overall quality of life for those who rely on ASL.

The advent of deep learning and computer vision technologies has opened new avenues for addressing these communication barriers. Recent advancements in object detection models, such as YOLO (You Only Look Once), have demonstrated remarkable capabilities in accurately identifying and localizing objects within images and video streams in real-time. The latest iteration, YOLOv10, offers enhanced performance and accuracy, making it an ideal candidate for developing an ASL letter detection model. This project leverages the strengths of YOLOv10 to create a real-time ASL letter detection model that uses a webcam to capture live video feed, detect ASL letters, and display them to facilitate communication between ASL users and non-users.

Our model is designed to process the video feed from a webcam, detect hand gestures corresponding to ASL letters, and provide visual feedback by annotating the detected letters on the video frames. The YOLOv10 model is trained on a custom dataset of ASL letters, which includes variations in hand positions, orientations, and lighting conditions to ensure robustness and accuracy. The use of a deep learning model allows the system to generalize well to new and unseen data, providing reliable detection in diverse real-world scenarios. By offering real-time detection and visual feedback, the model

aims to make ASL communication more accessible and effective.

Applications of this project extend beyond mere communication aid. In educational settings, it can serve as an interactive tool for teaching ASL to students and helping them practice their skills. For healthcare providers, it can assist in bridging the communication gap with deaf or hard-of-hearing patients, ensuring better understanding and care. Furthermore, the model can be integrated into customer service platforms, public information kiosks, and other interactive systems to provide inclusive services to ASL users. Our evaluation metrics demonstrate the model's effectiveness, with a mAP@50 of 94.20% and an overall accuracy of 92.86%. By harnessing the power of YOLOv10 and deep learning, this project not only enhances communication but also promotes inclusivity and accessibility in various domains of daily life.

II. LITERATURE SURVEY

The author of the paper [1] reviews several prior works on sign language recognition using various techniques. Traditional machine learning approaches, such as HOG with SVM and bag of visual words with SVM/KNN, have achieved accuracies of around 88-91% on alphabet datasets. Neural network models have been successful in recognizing ASL letters with up to 96.15% accuracy. Combining Leap Motion sensors with KNN/SVM has resulted in accuracies up to 79.83%. An end-to-end sensor-based method named "DeepSLR" was proposed for real-time detection with a larger vocabulary. Other studies have explored low-cost solutions, Indian Sign Language recognition, and benchmark datasets for evaluating classification models. Building on these efforts, the current study aims to leverage the latest advancements in YOLO object detection algorithms, specifically YOLOv5 and YOLOv8, to develop a custom CNN model for recognizing American Sign Language letters using publicly available ASL datasets.

The author of the paper [2] reviews prior works on gesture recognition from video sequences. Hidden Markov Models combined with Bayesian Network and Naive Bayes classifiers have been used for recognizing facial expressions from videos. PCA was explored for silhouette recognition and 3D modeling for human posture recognition, although intermediary gestures posed challenges. Other approaches segmented videos, extracted features, and classified them using distances like Euclidean and k-nearest neighbors. A pipeline for continuous Indian Sign Language recognition involving frame extraction, pre-processing, key frame selection, feature extraction using orientation histograms, and classification using distance measures was proposed. Building upon these methods, the current study aims to develop a vision-based application for American Sign Language recognition using deep learning techniques. This involves utilizing an Inception CNN to extract spatial features and a recurrent LSTM network to capture temporal features from video sequences of sign gestures.

The author of the paper [3] provides an overview of previous work in sign language recognition through computer vision. Techniques such as the Haar-cascade algorithm for detecting sign hand positions have been explored, along with datasets where the camera was positioned towards the signer. This indicates a thorough review of prior sign language datasets and related work involving vision methods and datasets. This groundwork laid the foundation for the proposed system, which combines hand position detection with a trained background model, aiming to advance sign language recognition through the integration of computer vision with machine learning.

The author of the paper [4] reviews a method to assist individuals with hearing impairments in human-to-human and human-to-robot sign language interactions using computer vision. Addressing challenges like background clutter and partial occlusion, the proposed SLR-YOLO network enhances feature extraction with an RFB module, utilizes BiFPN for feature fusion, and incorporates the Ghost module for a lighter network. Cutout data augmentation improves data generalization. The improved model achieves 90.6% accuracy on the American Sign Language Letters Dataset [12].

III. YOLOv10 Architecture: You Only Look Once Version 10

The YOLOv10 architecture consists of the following components:

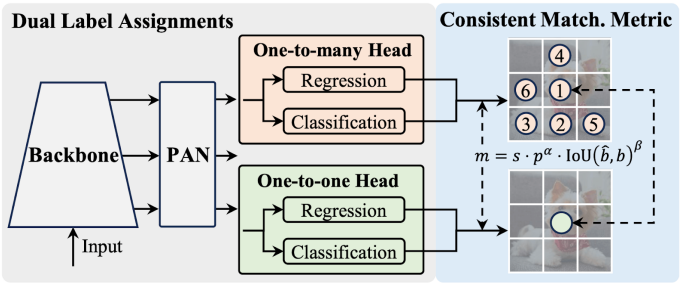

Fig. 1. YOLOv10 Architecture

A. Backbone

The backbone of YOLOv10 utilizes an advanced version of CSPNet (Cross Stage Partial Network). CSPNet enhances gradient flow and reduces computational redundancy, making feature extraction more efficient and effective.

B. Neck

YOLOv10 incorporates a neck designed with PAN (Path Aggregation Network) layers. These layers aggregate features from different scales to facilitate effective multi-scale feature fusion, improving the model's ability to capture and integrate information across scales.

C. One-to-Many Head

During training, YOLOv10 utilizes a One-to-Many Head. The head generates multiple predictions per object, providing rich supervisory signals to enhance learning accuracy. It allows the model to learn from diverse perspectives of object representation.

D. One-to-One Head

For inference, YOLOv10 switches to a streamlined One-to-One Head. This architecture generates a single best prediction per object, eliminating the need for Non-Maximum Suppression (NMS). This approach reduces latency and improves efficiency in real-time applications.

E. Objective Function

The objective function used in YOLOv10 can be expressed as follows:

$$m(\alpha, \beta) = s \cdot p^{\alpha} \cdot \text{IoU}(\hat{b}, b)^{\beta} \tag{1}$$

This equation combines the confidence score s, the probability p of the predicted class, and the Intersection over Union (IoU) between the predicted bounding box $\hat{b}$ and the ground truth bounding box b. The parameters α and β control the influence of the probability and IoU respectively. By optimizing this function, YOLOv10 aims to improve both the accuracy of class prediction and the precision of bounding box localization.

These architectural improvements enable YOLOv10 to achieve state-of-the-art performance in object detection tasks. The model excels in scenarios where real-time processing, accuracy, and efficiency are critical, making it suitable for a wide range of applications including ASL letter detection in live video streams.

IV. Proposed Work

A. Data Collection and Preprocessing

The development of an accurate ASL letter detection system requires a well-curated dataset. For this project, we collected a diverse set of images representing all the letters in the ASL alphabet, including variations in hand positioning, orientations, and lighting conditions. Images were labeled with corresponding ASL letter annotations.

To prepare the data for training, several preprocessing steps were implemented:

1) Image Augmentation: Techniques such as blurring, grayscaling, and contrast adjustment were applied to increase the diversity of the training set and improve the model's ability to generalize.

2) Normalization: Images were normalized to ensure consistency in the input data.

3) Label Encoding: ASL letters were encoded into numerical labels for model training purposes.

B. Model Training

The YOLOv10 model, known for its efficiency and accuracy in real-time object detection, was selected for this task. YOLOv10 (You Only Look Once version 10) builds on its predecessors by incorporating several enhancements, including improved anchor-free detection, better feature pyramid networks, and advanced augmentation techniques. The model is designed to detect multiple objects in an image with a single pass, making it highly efficient for real-time applications.

C. Model Training

The training process involved the following steps:

1) Model Initialization: The YOLOv10 model was initialized with pre-trained weights to leverage existing knowledge and accelerate the training process.

2) Dataset Loading: The custom ASL dataset was loaded, and data augmentation techniques were applied to enhance the training process.

3) Training Configuration: The model was configured to train over 100 epochs, optimizing the learning rate and momentum parameters to achieve the best performance.

4) Training Execution: The model was trained using the configured parameters, with continuous monitoring of validation accuracy to prevent overfitting.

D. Real-Time Detection

The core functionality of the system is its ability to perform real-time detection of ASL letters from a live video feed captured by a webcam. The following steps outline the real-time detection process:

1) Video Capture: The OpenCV library was used to capture frames from the webcam.

2) Frame Preprocessing: Each frame was preprocessed to convert it into a format suitable for the YOLOv10 model.

3) Model Inference: The preprocessed frames were fed into the YOLOv10 model to detect ASL letters. The model produced bounding boxes, class labels, and confidence scores for each detected letter.

4) Annotation: Detected letters were annotated on the video frames with bounding boxes and labels, providing immediate visual feedback.

5) Display: The annotated frames were displayed in real-time, allowing users to see the detected ASL letters instantly.

E. Dataset Details

During the training process, the dataset was thoroughly scanned and processed as follows:

- **Training Set:** 504 images of ASL letters.
- **Validation Set:** 144 images.

V. Experimental Results

The overall performance of the ASL letter detection system was evaluated using the following metrics: precision, recall, mean Average Precision (mAP) at IoU 0.50, mAP at IoU 0.50-0.95, and accuracy. The results are summarized in Table I.

Precision refers to the proportion of true positive detections among all positive detections made by the model. A precision of 87.60% indicates that 87.60% of the letters detected by the model were correct. Recall measures the proportion of actual positives that were correctly identified by the model, with a recall of 87.00% meaning that the model identified 87.00% of all actual letters correctly.

The mean Average Precision (mAP) at IoU 0.50 and IoU 0.50-0.95 provides a comprehensive evaluation of the model's performance. mAP@50 is calculated at a single IoU threshold of 0.50, yielding 94.20%, which indicates high precision and recall at this threshold. mAP@50-95, which averages precision over multiple IoU thresholds from 0.50 to 0.95, is 90.20%,

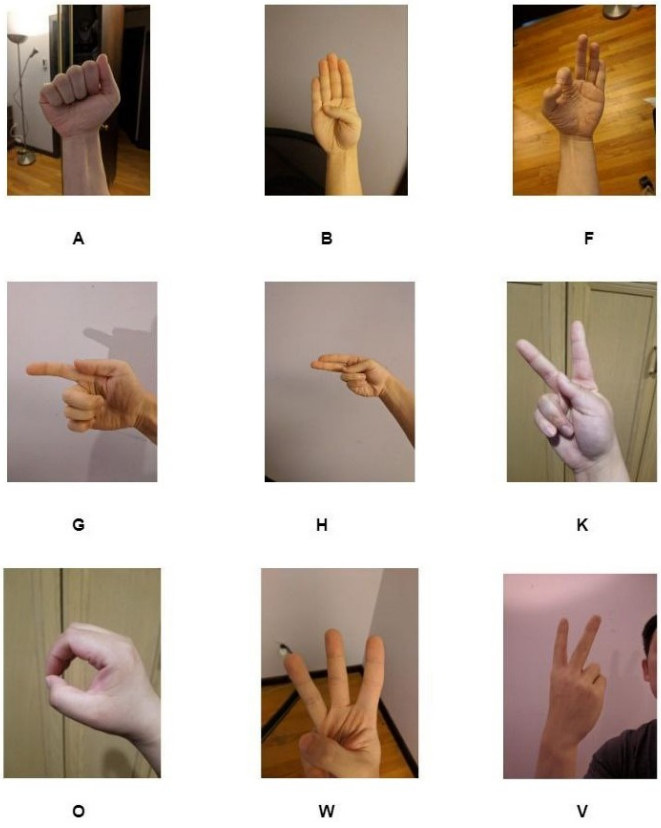

Fig. 2. Dataset.

TABLE I
OVERALL PERFORMANCE METRICS

Metric	Value
Precision	87.60%
Recall	87.00%
mAP@50	94.20%
mAP@50-95	90.20%
Accuracy	92.86%

reflecting the model's ability to maintain high accuracy across different levels of localization precision. Finally, the overall accuracy of the model is 92.86%, indicating that 92.86% of the total predictions were correct.

Figure 3 illustrates the training and validation loss curves. The loss function, which combines confidence loss, classification loss, and bounding box regression loss, was minimized throughout the training process. A decreasing trend in both training and validation loss curves suggests that the model is learning effectively without overfitting.

Figure 4 shows the confusion matrix for the ASL letter detection system. The confusion matrix provides a detailed

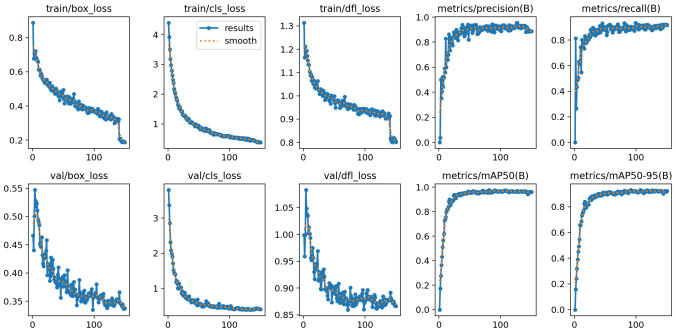

Fig. 3. Training and Validation Loss.

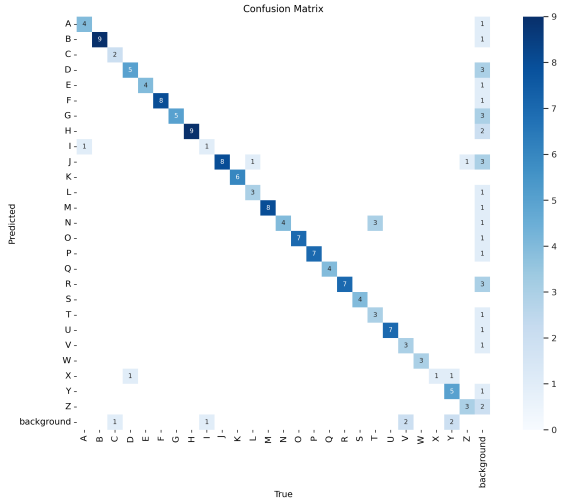

Fig. 4. Confusion Matrix.

breakdown of the model's performance, indicating how often letters are correctly identified and where misclassifications occur.

The following figures show examples of the detected ASL letters for various letters in the alphabet, demonstrating the system's capability to accurately recognize and label hand signs.

These results demonstrate the robustness and accuracy of the YOLOv10-based ASL letter detection system, highlighting its potential for real-time applications in translating sign language.

VI. CONCLUSION

Our system uses the YOLOv10 model for real-time ASL letter detection, trained on a diverse dataset to ensure accuracy and robustness. Video frames from a webcam are preprocessed and fed into YOLOv10, which detects and annotates ASL letters. OpenCV handles video capture and display, while PyTorch manages model inference, ensuring high accuracy and low latency. This system can be used as an educational tool for teaching ASL, aiding healthcare providers in communicating with deaf or hard-of-hearing patients, and enhancing customer service and public information kiosks for ASL users.

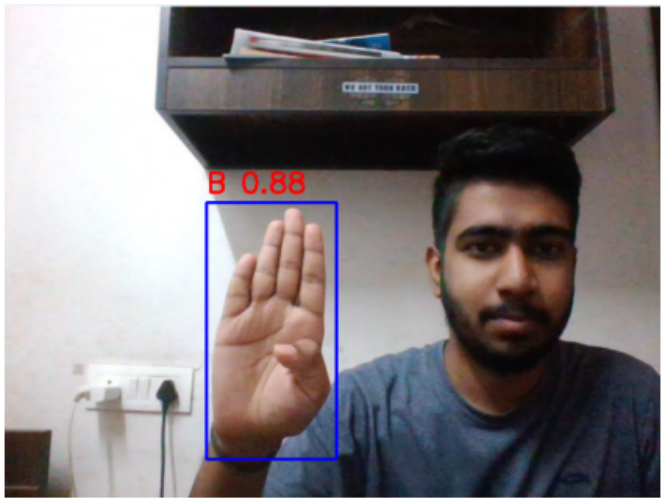

Fig. 5. Sign Language for Letter B.

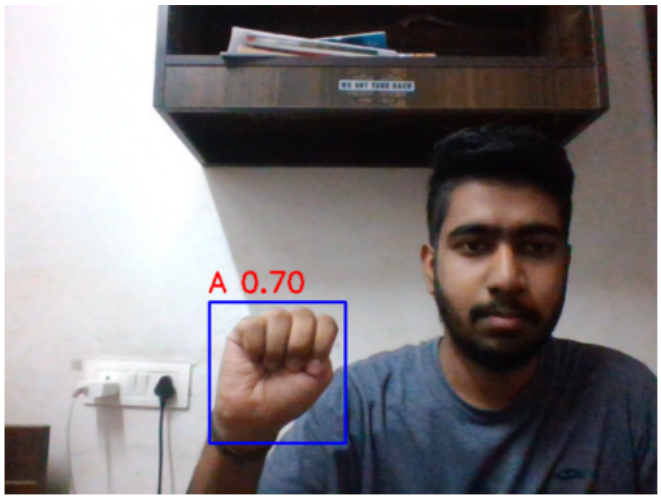

Fig. 6. Sign Language for Letter A.

By leveraging YOLOv10, our project promotes inclusivity and accessibility, highlighting the transformative potential of assistive technologies.

VII. FUTURE SCOPE

Future work will prioritize expanding the size and diversity of the ASL dataset. A larger and more varied dataset will encompass a broader range of hand shapes, orientations, lighting conditions, and backgrounds, which will enhance the model's robustness and accuracy. By incorporating more examples of each ASL letter and diversifying the training data with different skin tones, hand sizes, and environmental factors, the model can generalize better to real-world scenarios. This expansion is crucial for improving the reliability of the system in recognizing ASL letters across different users and contexts, ultimately making the detection more inclusive and precise.

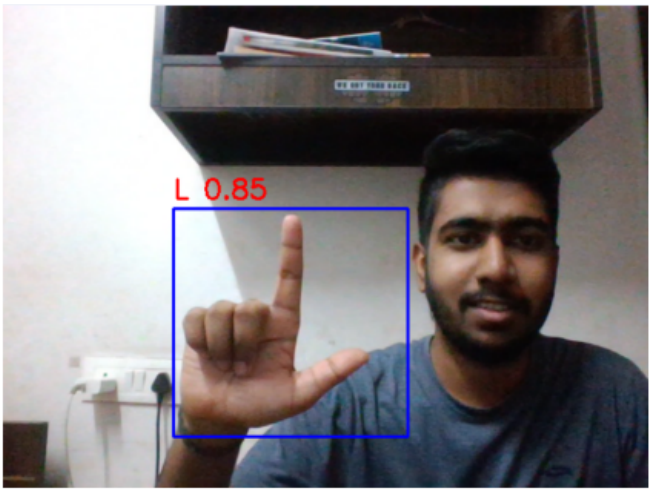

Fig. 7. Sign Language for Letter L.

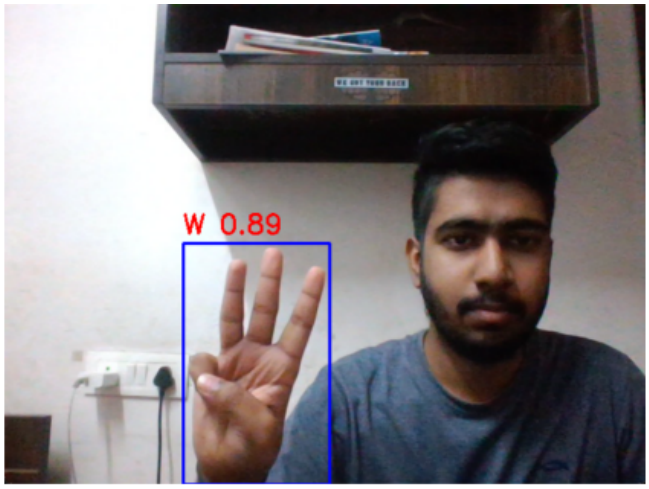

Fig. 8. Sign Language for Letter W.

REFERENCES

[1] Shaji, Aleena, and Ms Resija PR. "A REVIEW ON SIGN LANGUAGE RECOGNITION."

[2] Bantupalli, Kshitij, and Ying Xie. "American sign language recognition using deep learning and computer vision." In 2018 IEEE International Conference on Big Data (Big Data), pp. 4896-4899. IEEE, 2018.

[3] Savur, Celal, and Ferat Sahin. "Real-time american sign language recognition system using surface emg signal." In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pp. 497-502. IEEE, 2015.

[4] Nareshkumar, M. Daniel, and B. Jaison. "A Light-Weight Deep Learning-Based Architecture for Sign Language Classification." Intelligent Automation Soft Computing 35, no. 3 (2023).

[5] Triwijoyo, Bambang Krismono, Lalu Yuda Rahmani Karnaen, and Ahmat Adil. "Deep learning approach for sign language recognition." JITEKI: Jurnal Ilmiah Teknik Elektro Komputer dan Informatika 9, no. 1 (2023).

[6] Sahoo, Ashok K., Gouri Sankar Mishra, and Kiran Kumar Ravulakollu. "Sign language recognition: State of the art." ARPN Journal of Engineering and Applied Sciences 9, no. 2 (2014): 116-134.

[7] Pathan, Refat Khan, Munmun Biswas, Suraiya Yasmin, Mayeen Uddin Khandaker, Mohammad Salman, and Ahmed AF Youssef. "Sign

language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network." Scientific Reports 13, no. 1 (2023): 16975.

[8] Srivastava, Sharvani, Amisha Gangwar, Richa Mishra, and Sudhakar Singh. "Sign language recognition system using TensorFlow object detection API." In International conference on advanced network technologies and intelligent computing, pp. 634-646. Cham: Springer International Publishing, 2021.

[9] Kothadiya, Deep, Chintan Bhatt, Krenil Sapariya, Kevin Patel, Ana-Belén Gil-González, and Juan M. Corchado. "Deepsign: Sign language detection and recognition using deep learning." Electronics 11, no. 11 (2022): 1780.

[10] Kumar, Anup, Karun Thankachan, and Mevin M. Dominic. "Sign language recognition." In 2016 3rd international conference on recent advances in information technology (RAIT), pp. 422-428. IEEE, 2016.

[11] Purnomo, Hindriyanto, Christine Dewi, Budhi Kristanto, Kristoko Hartomo, and Siti Hashim. "Utilizing the YOLOv10 Model for Accurate Hand Gesture Recognition with Complex Background." Available at SSRN 4777516.

[12] Jia, Wanjun Li, Changyong. (2023). SLR-YOLO: An improved YOLOv10 network for real-time sign language recognition. Journal of Intelligent Fuzzy Systems. 46. 1-18. 10.3233/JIFS-235132.