Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Hybrid YOLOv8-seg Model for Hand Gesture Segmentation

Avadhoot R. Telepatil[1], Vaddin J S.[2]

[1]PhD Scholar, DKTE'S TEI, Ichalkaranji, Shivaji University, Kolhapur, Maharashtra, India.

[2]Professor, Former HOD Electrical & Professor Electronics (PG), DKTE's Textile and Engineering Institute, Maharashtra, India.

E-mail: [1]avadhootrtelepatil@gmail.com, [2]vjayashree@dkte.ac.in

Abstract

Parkinson's Disease (PD) is a neurological disorder that causes patients with Parkinson's Disease (PPD) to have difficulty with body balancing. Thus, PPD rely on caretakers to fulfill their daily needs. Vision-based assistive systems could be useful for PPD to communicate with caretakers. The work presented here is a hybrid YOLOv8n-seg framework with DETR. In this framework, the traditional YOLOv8n-seg model's head is replaced with DETR as the head for hand gesture (HG) segmentation for PPD. Since no public dataset exists for PPD gestures, a dataset of 4,583 raw hand gesture images was collected with a webcam under realistic home and clinical environments (such as poor light, cluttered background, and motion blur) and expanded via augmentation to 11,230 gestures. This dataset was divided into an 80% training set, a 15% validation set, and a 5% testing set with 9 classes (e.g., hungry, attention, call, toilet) to ensure robust evaluation. The baseline YOLOv8n-seg model and Transformer-based variant, DETR (DEtection TRansformer), were tested on the custom PPD hand gesture dataset. Compared to the baseline YOLOv8n-seg, the implemented hybrid model achieved superior performance across all evaluation metrics, with ~1% improvement in precision (99% vs. 98%), recall (97% vs. 96%), F1 score (98% vs. 97%), and dice score (98% vs. 97%), with almost the same mAP@50 (97% vs. 97% for all), while improving inference speed by +3.0% (55.1 FPS vs. 53.5 FPS). On the same custom dataset, the conventional U-Net achieved 88% precision, 92% recall, and a 0.9 dice score, whereas the proposed hybrid model reached 99% precision, 98% recall, and a 0.98 dice score. This confirms the superior performance of the hybrid model over the conventional U-Net architecture for HG segmentation. The Raspberry Pi 4B is used as an edge device for HGR of PPD. These enhancements demonstrate that the hybrid approach achieves both higher accuracy and faster real-time performance, which is useful for assistive systems deployment on the embedded edge device. To our knowledge, this is the first work combining YOLOv8n-seg with a DETR head for PPD hand gesture segmentation.

Keywords: Hand Gesture Segmentation, Parkinson's Disease Patient, Transformer Network, YOLOv8n-Seg, Assistive Technology.

1. Introduction

Parkinson's Disease (PD) is a neurodegenerative condition that attacks motor control, inducing symptoms of tremor, rigidity, and slowness [1]. Such difficulties render activities of daily living, such as walking, dressing, or eating, troublesome for PD patients. Recent

DOI: https://doi.org/10.36548/jiip.2025.3.017
Received: 17.07.2025, received in revised form: 21.08.2025, accepted: 06.09.2025, published: 19.09.2025

technological advancements based on wearable technology are emerging as valuable aids. These technologies assist in the real-time monitoring of symptoms, offering patients and clinicians alike valuable information for improved management. However, PD patients have to wear the wearable sensors or devices on their bodies, which can be cumbersome. Image-based sign language detection and recognition is a new assistive technology that uses computer vision and machine learning to translate hand movements into text or speech. For patients with Parkinson's Disease (PD) and other speech disorders, the technology provides an easy, non-invasive way to navigate around motor obstacles. In contrast to wearable systems, which can be obtrusive for PD patients, image-based systems utilize cameras to capture gestures in real time, facilitating effortless communication. This refers [2] to the impact of non-intrusive, machine learning-based wearable systems on controlling motor symptoms such as freezing of gait. This is supplemented by Deb R et al. [3], where researchers have identified four target areas of research such as diagnosis, monitoring, response to treatment, and rehabilitation and further recommend multi-modal & user-centered systems. Voice-assisted technologies (VAT) have also emerged as useful for PD patients who have speech impairments. Duffy et al. [4] present positive user feedback but require additional clinical trials.

Experiments conducted by Kour et al. [5] and Dineskkumar et al. [6] confirm the potential of AI-enabled, sensor-based gait analysis in detecting tremors and diagnosing early, providing low-cost, non-surgical solutions. In addition, Deb et al. [7] demonstrate the growing emphasis on wearable technologies for tremor and gait analysis through machine learning. Although these developments have been accomplished, issues of ideal sensor placement, limited datasets, and model precision must still be resolved. Therefore, the integration of sign language via hand gesture recognition presents a novel contribution to existing assistive technologies for PD, especially for patients with communication issues. It aids in non-verbal communication, facilitating greater independence and daily interaction for patients. YOLOv8 has been widely applied to most instance segmentation tasks in various areas, demonstrating its superior performance and universal applicability in plant leaf segmentation, The YOLOv8-seg model was improved by adding the Ghost and BiFPN modules by Wang et al. [8], significantly enhancing the segmentation of small leaves. This innovation resulted in an 86.4% Dice score in the CVPPP challenge, outperforming state-of-the-art approaches. Work presented by Alsuwaylimi et al. [9] employed improved YOLOv8n-Seg and YOLOv8s-Seg models for real-time underwater garbage detection with high accuracy and speed, rendering them viable for marine conservation. The YOLOv8m-Pp coral instance segmentation model by Hassanudin et al. [10], achieving 96.7% accuracy and 98.2% mAP@50, showing the model's effectiveness in ecological monitoring. Comparison between YOLOv8 and YOLOv5 by Casas et al. [11] for corrosion detection on metallic surfaces concludes that YOLOv8 is superior in segmentation accuracy, speed, and robustness.

Improved YOLOv8-seg is suggested by Bai et al. [12] for monitoring construction sites with UAVs, adding more modules to enhance feature extraction and context understanding, leading to improved performance and efficiency. Within agriculture, Sapkota et al. [13] demonstrated the superiority of YOLOv8 over Mask R-CNN in orchard segmentation with better accuracy and faster inference times, proving it highly suitable for real-time agricultural automation tasks like robotic harvesting. YOLOv8 has been a very valuable instrument in a wide range of segmentation tasks in general, and it has performed well in both accuracy and speed. Transformer-based models have greatly improved visual segmentation and serve as strong substitutes for conventional convolutional neural networks (CNNs). Work accomplished by Li X et al. in [14] provides a general overview of transformer-based segmentation methods, focusing on unified meta-architectures and task-specialized applications in areas such as

medical imaging and 3D point cloud segmentation. DETR (DEtection TRansformer), developed by N.

Carion [15] transformed object detection with its transformer-based approach, whereas issues with small object localization and convergence speed have led to further improvements. Tian et al. [16] generalized DETR to layered clothing segmentation and achieved state-of-the-art performance on the Fashionpedia dataset. Zhang et al. [17] also proposed an improved Mask Transformer with better cervical cell segmentation accuracy through innovations such as dynamic anchor boxes and noised ground truth embeddings. Other transformer-based models, like UW-DETR by Jiang et al. [18], also achieved success in underwater segmentation. H Kadi et al. in [19] employed DETR to illustrate its application in medical imaging, where it detects and segments teeth in panoramic radiographs. These successes highlight the effectiveness and versatility of transformer-based models across different fields. Wang et al. [20] proposed a hybrid segmentation method through the integration of YOLOv5 and the Segment Anything Model (SAM) for precise detection of water transport patterns in textiles with superior performance to state-of-the-art alternatives. Tang et al. [21] extended the application of transformers by designing SFED-Former as a superior version of DETR for video lane detection, benefiting from improved training efficiency and detection accuracy for autonomous driving. The ability to use transformer models on a variety of challenging tasks, ranging from biomedical image segmentation to lane detection based on video, shows their promise for expert and precise segmentation.

YOLOv8, with its high precision and real-time inference speed, is becoming a top contender for instance segmentation in agricultural monitoring, marine conservation monitoring, and industrial inspection applications. Its real-time application efficiency makes it a viable candidate for feature applications like robotic farming, underwater trash detection, and construction site monitoring by R. Bai et al. [22]. State-of-the-art segmentation approaches such as YOLOv8n-seg and U-Net have high segmentation accuracy at high computational expense. The proposed YOLOv8n-seg fused with DETR as the head consolidates the real-time, lightweight inference properties of YOLOv8n-seg with the global attention mechanism of DETR, establishing it as the very first framework for PPD hand gesture segmentation. This approach enabling robust, edge-deployable solutions in assistive healthcare systems for PPD. In particular, we confirmed its feasibility on Raspberry Pi 4B as a low-cost embedded platform. This confirms that the suggested framework is not only suitable but also feasible for real-time assistive applications for PPD in real-world environments.

2. Objective

The main objective of the proposed research is to design an effective and precise hand gesture segmentation framework specifically for PPD with the ability to perform in real time and with high accuracy in assistive technology applications. To meet this purpose, the following specific objectives are defined:

1. To create and build a custom dataset of varied hand gestures for PPD, in response to the unavailability of publicly available PDP HG datasets.

2. To use the baseline YOLOv8n-Seg model for the segmentation of hand gestures, providing a strong foundation and reference point for measuring improvements introduced by sophisticated segmentation heads.

3. To extend the YOLOv8n-Seg architecture by adding a lightweight DETR (DEtection TRansformer) head, with the goal of enhancing segmentation performance without loss of real-time inference efficiency.

4. To compare the baseline and Transformer-augmented models based on quantitative metrics to evaluate the performance and robustness of the proposed solution.

3. Methodology

The YOLOv8n-seg model, while effective, is hampered in its ability to detect the subtle, tremor-impacted hand movements routinely demonstrated by PPD because it lacks global context awareness and attention. The proposed methodology aims to improve the YOLOv8n-Seg model's performance in hand gesture segmentation through the integration of Transformer-based DETR architectures as heads. These changes overcome the shortcomings of the original YOLOv8n-Seg model by enhancing its segmentation accuracy and efficiency, while preserving the real-time inference capabilities necessary for practical use. The performance of the developed models was tested on a custom hand gesture segmentation dataset prepared by us on Google Colab, which had an L4 GPU. The DETR Transformer-improved YOLOv8n-Seg was trained and tested under the same conditions to make a valid comparison. The workflow is as illustrated in Figure 1.

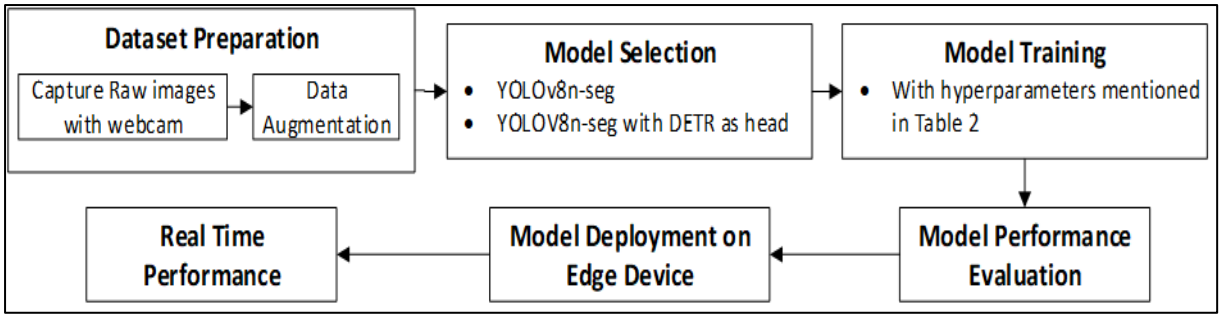


Figure 1. Workflow of Implemented Work

Initially, the custom hand gesture dataset was prepared. The hand gesture images were captured with a USB webcam under real-world environments, such as cluttered backgrounds, varying light conditions, and variations in hand orientation. The captured images were not enough to train and validate the deep learning model; hence, the dataset was enhanced to 11,230 sample images using data augmentation techniques, such as rotation, scaling, and blur. The dataset was then divided into 80% training, 15% validation, and 5% testing sets with 9 classes (e.g., hungry, attention, call, toilet) to ensure robust evaluation of the YOLOv8n-seg and YOLOv8n-seg integrated with the DETR head. Both selected models were trained on the custom dataset. The model performance was then analyzed using the evaluation parameters mentioned in section 3.5.

The hybrid YOLOv8n-seg integrated with the DETR head model's performance was found to be good compared to the YOLOv8n-seg after analyzing the evaluation parameters. Thus, the hybrid model was deployed on the Raspberry Pi 4B as an embedded edge device. Finally, the designed framework was tested in real-world, real-time environmental conditions. The detailed results are presented further.

3.1 Dataset Preparation

The performance of any vision based system is highly dependent on the quality and diversity of the dataset. Most publicly available hand gesture datasets have been captured from healthy people under controlled lighting and background conditions with limited hand gesture variability. Such datasets do not adequately represent the muscle movement difficulties experienced by individuals with Parkinson's disease (PPD). This limits the use of publicly available datasets in vision-based assistive systems. A publicly available hand gesture dataset for PPD is not available. We created a custom hand gesture dataset that includes gestures performed by healthy as well as elderly people exhibiting muscle stiffness, one of the key symptoms of PPD. This dataset consists of 4,583 original hand gesture images, captured using standard USB webcam in a realistic and cluttered environment including variations in lighting, gesture orientation, hand size, skin tone and background complexity. To improve model generalization and prevent overfitting, data augmentation techniques viz.; rotation, scaling, and blurring were applied. This increased the dataset size to a total of 11,230 images. The dataset covers 9 gesture classes representing essential daily needs for individuals with PPD (e.g., water, hunger, natural call, toilet etc.). Care was taken to preserve class balance during augmentation so that each class remained equally represented. For model development, the dataset was divided into training, validation and testing sets using an 80%,15%, and 5% ratio. This ensures sufficient data samples for model training, validation and unbiased testing. An overview of the dataset is presented in Table 1 where each row represents a hand gesture assigned with the message to be communicated by individuals with PPD, along with the associated class.

Table 1. Hand Gestures used by PDP to Express Need

Sr. No.	Posture	Meaning		Sr. No.	Posture	Meaning
1		Indicates thirst [class : water]		6		Indicates a desire to move [class : movement]
2		Signals hunger [class : hungry]		7		Signal to turn ON/OFF device [class : control]
3		Signifies a natural call [class : natural call]		8		Represents agree [class : agree]

4		Urge to use the toilet [class : toilet]		9		express happiness [class : happy]
5		Seeks for attention [class : attention]				

3.2 YOLO8n-seg Model

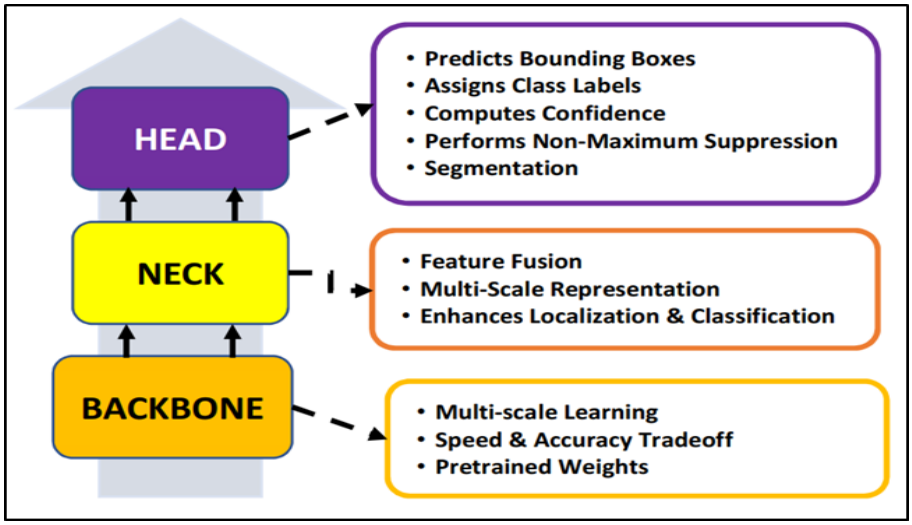


Figure 2. Outline of YOLO Model Architecture

YOLOv8 was released by Ultralytics in January 2023. One of its variants, YOLOv8n is a lightweight, real-time model for object detection and instance segmentation. This model has three sections, backbone, neck and head. The outline of the YOLO model is as shown in Figure 2. It uses a CSPDarknet backbone with a deeper and better efficient Cross Stage Partial (CSP) design. This enables stronger gradient flow and more feature representation, improving the detection of fine-grained details such as hand edges and contours. A PANet neck for multi-scale feature fusion. The PANet combines features from different layers and strengthens the flow of localization signals from shadow layers. This ensures better boundary localization under cluttered and noisy background. Finally, there is an anchor-free head that predicts classes, boxes, and segmentation masks in a single-stage pipeline. Its mask branch enables efficient instance-level segmentation with minimal computational overhead. In comparison with YOLOv5, the YOLOv8-seg backbone improves feature extraction. The YOLOv5 head is based on the anchor design whereas YOLOv8-seg uses an anchor free detection head. This eliminates the need for manual anchor tuning and improves the localization of irregular and fine-grained hand shapes. While YOLOv5 lacks native segmentation, YOLOv8-seg integrates a mask prediction branch directly into the backbone neck pipeline ensuring that the learned features contribute simultaneously to detection and segmentation, helping to enhance overall accuracy and efficiency.

3.3 DETR

DETR, through the capability of transformers, provides a very efficient and effective method for segmentation tasks. Object queries for segmentation, along with its end-to-end training system, make DETR a robust substitute for conventional CNN-based segmentation models. Semantic segmentation classifies every pixel in an image into groups but cannot distinguish one from another if two are from the same category. Instance segmentation continues by separating specific instances of objects that belong to the same group and assigning every instance a separate label. With direct pixel-level mask prediction and bounding box refinement, DETR is capable of performing both semantic and instance segmentation, delivering useful results for numerous applications, including hand gesture segmentation and medical image analysis. In our presented work, the traditional segmentation head of the YOLOv8n-seg model is substituted with a transformer-based DETR head. DETR eliminates the need for handcrafted components such as anchor boxes, non-maximum suppression (NMS) and region proposal networks, all while providing a unified architecture based on the Transformer model. The core innovation lies in leveraging self-attention mechanisms for global context reasoning, while is essential for capturing subtle hand gesture variations associated with Parkinsonian symptoms. The hybrid model is an integration of the real-time inference of YOLOv8n with the pixel-level segmentation precision of DETR. End-to-end training with object queries and direct prediction of masks is facilitated through the integration. The consequence of this customized YOLOv8n-DETR model shows enhanced segmentation performance (as shown in section 4) for applications such as hand gesture segmentation.

3.4 Hyperparameter Settings

To ensure optimal training performance and generalization capability of the segmentation model, some of the important hyperparameters were accurately chosen and optimized. The details of the set hyperparameters are outlined in Table 2.

Table 2. Hyperparameters used for Training the Models

Hyperparameter	Value	Justification
Epochs	50	The model stabilized after ~40 epochs & 50 ensures optimal training without overfitting
Image size	640 x 640	The smaller size led to information loss of fine hand details whereas the larger image size increases the computation efficiency.
Batch	16	Selected to avoid memory overflow for L4 GPU
Optimizer	RMSProp	RMSProp showed smoother convergence and lower validation loss for our custom dataset [1]
Learning rate	0.001	Higher rate caused divergence while lower rate slowed convergence
Weight_decays	1.00e-04	Used for regularization to reduce overfitting on given dataset size

**In our previous work [1], we conducted a comparative study of five different optimizers viz.; SGD, Adagrad, Adadelta, RMSprop and Adam on custom hand gesture dataset. Among these, RMSprop achieved highest performance. Therefore, it was chosen the optimizer for model training in this study*

3.5 Evaluation Parameters

To correctly measure the instance segmentation model's performance, Precision, Recall, F1 Score, Dice Score, and Mean Average Precision (mAP) at two Intersection-over-Union (IoU) levels: mAP@50 and mAP@75 are considered as evaluation parameters. These measurements give an overall score across both detection accuracy and mask quality.

Precision measures the proportion of correctly segmented pixels (true positives) among all pixels predicted as positive:

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

Recall quantifies the proportion of correctly segmented pixels among all ground truth positive pixels:

$$Recall = \frac{TP}{TP+FP} \tag{2}$$

Here,

TP: A pixel belongs to an object that is correctly predicted as part of the object.

FP: A pixel does not belong to an object that is predicted as part of the object

TN: A pixel does not belong to an object that is correctly predicted as not a part of the object

FN: A pixel that belongs to the object but is predicted as not belonging to the object.

F1 Score is the harmonic mean of Precision and Recall and provides a balanced measure of segmentation accuracy:

$$F1\ score = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \tag{3}$$

Dice Score evaluates the spatial overlap between the predicted and ground truth segmentation masks. It is particularly useful in imbalanced datasets, such as those found in medical image segmentation:

$$Dice\ Score = \frac{2\ |P \cap G|}{|P|+|G|} \tag{4}$$

Here, P and G represent the set of predicted and ground truth pixels, respectively.

3.5.5 Mean Average Precision (mAP) at IoU thresholds of 0.50 (mAP@50) and 0.75 (mAP@75), following the COCO-style evaluation. The mAP metric assesses both the correctness of the segmentation mask and its localization with respect to the ground truth. For each class, Average Precision (AP) is computed based on the area under the Precision-Recall curve. The mean is then calculated across all classes:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (5)$$

Where N is number of classes in dataset.

3.6 Framework Design and Implementation

The real time framework setup for hand gesture capture and segmentation is shown in Figure 3(a). The Raspberry Pi 4B module is used as a main processing component in the designed framework. The hybrid YOLOv8n-seg model integrated with the DETR is deployed on the Raspberry Pi module. The USB camera is plugged into the Raspberry Pi 4B via USB 2.0, the system's CPU. The process begins with live recording of the video, which is done with the help of the USB webcam. Figure 3(b) shows the real-time webcam feeds to the system.

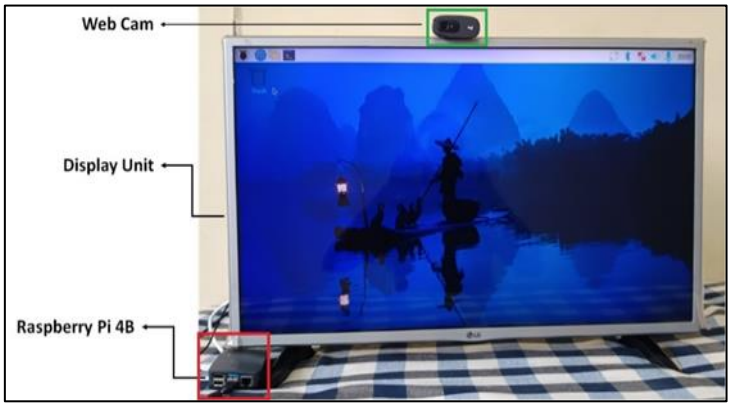


Figure 3 (a) Real Time Setup for HGR Framework

Figure 3 (b) Realtime Feed on Display unit by USB Camera

The detailed performance of the designed framework with visual results is discussed in section 4. The detailed result of the model's performance on custom dataset is detailed in section next.

4. Experimental Results and Discussion

This section explains training and validation performance plots and quantitative evaluation metrics for all model variants. Important indicators like Precision, Recall, F1 Score,

Dice Score, and mAP are compared. Lastly, visual segmentation results show the improvements made using Transformer-based heads compared to the baseline YOLOv8n-seg model.

Figure 4 indicates YOLOv8n-seg training and validation loss (box, cls, DICE, and segmentation) converging nicely over 50 epochs, reflecting stable learning and generalization. The measured metrics Precision, Recall, mAP@50, and mAP@75 - steadily increase and settle at high values, establishing the robustness of the baseline model and rendering it a good comparator against Transformer-based variants.

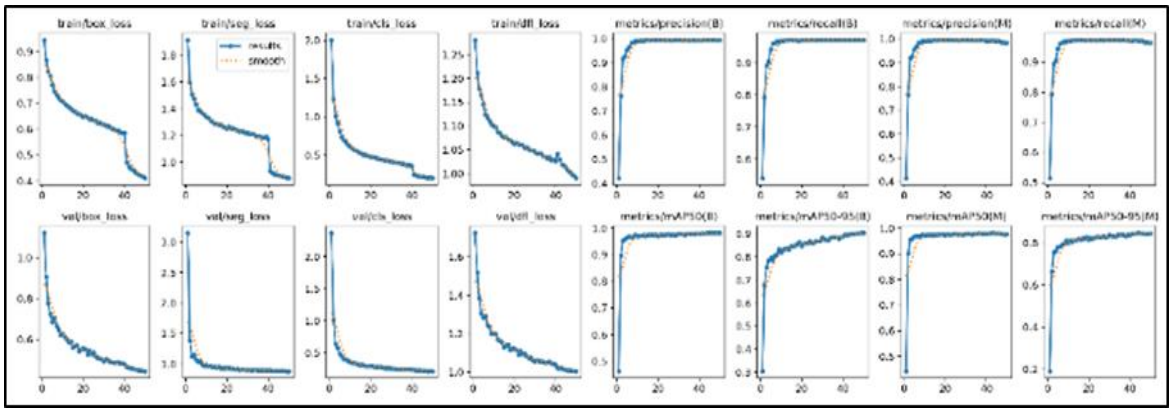


Figure 4. Training and Validation Performance Metrics for Baseline YOLOv8n-Seg Model

The training and validation performance for YOLOv8n-Seg with DETR transformer head is as shown in Figure 5.

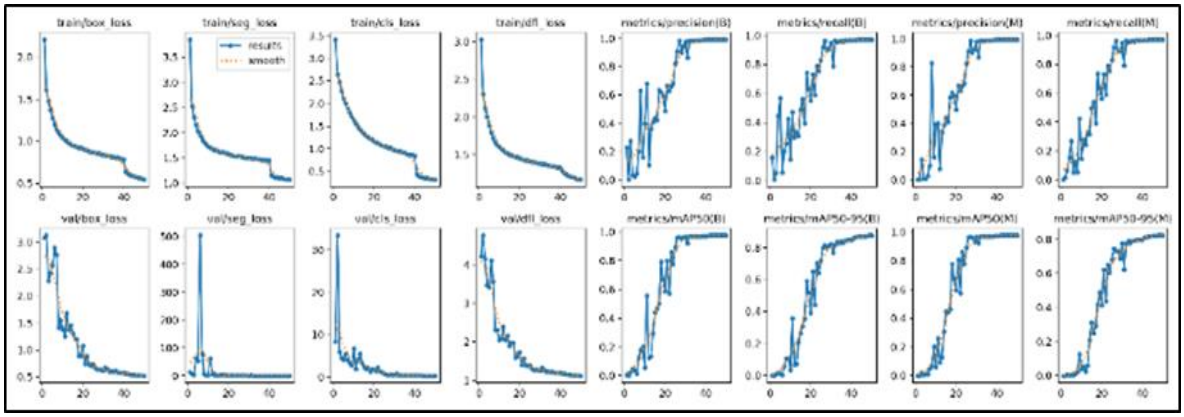


Figure 5. Performance Metrics of YOLOv8n-Seg with DETR Transformer-Based Head

Over 50 epochs, the model with a DETR Transformer-head exhibits smooth convergence in all the loss metrics box, classification, DICE segmentation, and DFL reflecting stable and efficient learning. Major evaluation metric viz Precision, Recall, mAP@50, and mAP@50-95 are all consistently improving and achieving high values for background and mask classes. This is evidence of the DETR head's capability to represent spatial relationships and serves as an effective baseline for attention-based segmentation methods. Table 3 shows the performance measures of the YOLOv8n segmentation model trained and tested on a custom hand gesture segmentation dataset.

Table 3. Performance metrics of YOLOv8n-Seg for Segmentation Task

Metric	Train	Validation	Test
Precision	98%	98%	98%
Recall	96%	96%	96%
F1 Score	97%	97%	97%
Dice Score	0.97	0.97	0.97
mAP@50	97%	97%	97%
mAP@75	97%	97%	97%
Time (in sec)	7389.19	15.28	10.5

The model shows excellent accuracy in training, validation, and test sets, with precision, recall, F1 score, Dice score, and mAP (at IoU thresholds of 0.5 and 0.75) all averaging 97% - 98%. Training took around 7389.19 seconds, while validation and test times were much lower at 15.28 sec and 10.5 sec respectively, showing the model's real-time applicability. Table 4 shows the performance results of the adapted YOLOv8n-Seg model that has a DETR-based Transformer head, tested on a self-created hand gesture segmentation dataset. The model had a fixed precision of 99% during training, validation, and testing. Recall values were 97% for training and validation, and 98% for the test set. The equivalent F1 score is 98% and the Dice score is 0.98, indicating a good trade-off between precision and recall. The mAP@50 and mAP@75 scores trended similarly at 97% (train and validation) and 98% (test), indicating strong segmentation performance. Training time took around 7384.96 seconds, with validation time and test time being 14.82 seconds and 10.2 seconds respectively, indicating the readiness of the model for near real-time application with enhanced accuracy.

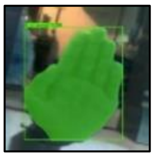
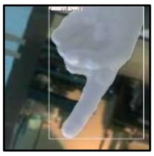
Table 4. Performance metrics of YOLOv8n-Seg with DETR head for Segmentation Task

Metric	Train	Validation	Test
Precision	99%	99%	99%
Recall	97%	97%	98%
F1 Score	98%	98%	98%
Dice Score	0.98	0.98	0.98
mAP@50	97%	97%	98%
mAP@75	97%	97%	98%
Time (in sec)	7384.96	14.82	10.2

Table 5 shows some sample segmentation outputs by the baseline YOLOv8n-seg model and its Transformer-enriched variation DETR segmentation heads. The baseline model performs reasonably but tends to fail to accurately mark certain hand boundaries in some cases. By contrast, our custom model generates much sharper and more complete masks. The DETR-integrated model shows robustness over diverse backgrounds and lighting. These visual findings support the quantitative ones and demonstrate that the use of Transformer-based

segmentation heads with YOLOv8n-seg greatly enhances the model's ability to segment hand gestures accurately in Parkinson's Disease patients.

Table 5. Qualitative Comparison of Hand Gesture Segmentation Models on PDP Data

Test Image	YOLO8n-Seg Model	Custom YOLO8n-seg (DETR head)	Test Image	YOLO8n-Seg Model	Custom YOLO8n-seg (DETR head)
					
					
					
					

The comparative performance plot shown in Figure 6(a) & 6(b) compares two models, YOLO8n-seg and a Custom YOLO model, on training, validation, and test datasets based on various performance metrics. The metrics used are Precision, Recall, F1 Score, Dice Score, mAP@50, and mAP@75. The comparison indicates that the DETR Transformer-enhanced YOLOv8n-Seg model provides the optimum overall performance. It has the highest precision of 99% in all phases, complemented by excellent recall (0.97) and F1 and Dice scores of 0.98 Additionally, it logs the lowest training time (7384.96 sec) and inference times, making it both accurate and efficient. These findings demonstrate that the DETR Transformer head provides a better trade-off between segmentation quality and real-time usability than the baseline approaches. Both models show good overall performance, with the Custom YOLO model performing slightly better than the YOLO8n-seg model on most metrics. In particular, the Custom YOLO model registers the best Precision (99%), F1 Score (98%), and Dice Score (0.98) on the test dataset, signifying its excellent capacity to accurately detect and segment pertinent features. In addition, the Custom YOLO model demonstrates better generalization ability as indicated by its higher mAP@50 and mAP@75 scores (98%) when tested. The

outcome indicates that the Custom YOLO model is stronger and more trustworthy in real-world application situations, especially in the segmentation process.

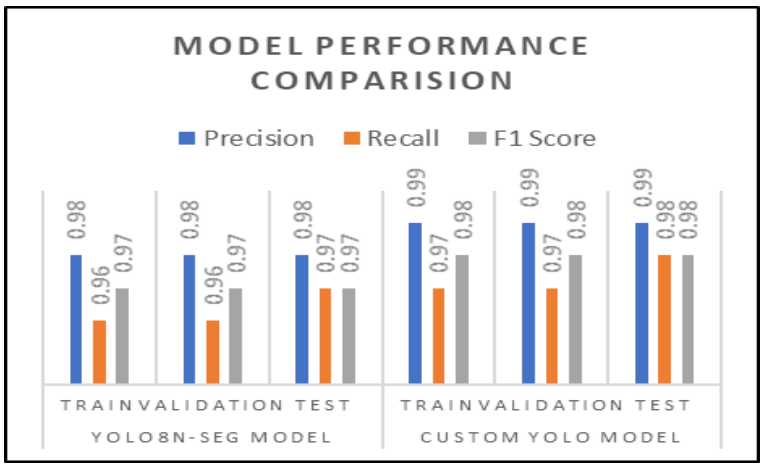


Figure 6 (a). Model Performance Comparison with Precision, Recall and F1 Score

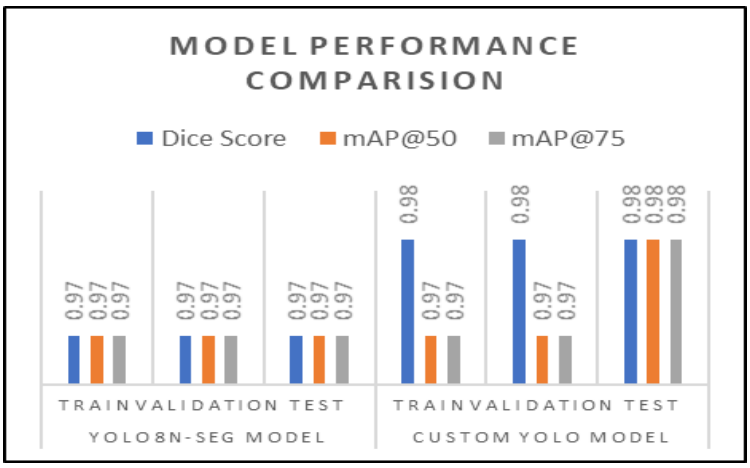


Figure 6 (b). Model Performance Comparison with Dice Score, mAP@50, mAP@75

Further the model performance is compared with the conventional U-Net architecture and the results are represented in Table 7. U-Net performs significantly lower (Precision: 88%, Recall: 92%, F1 score: 89%, Dice score 0.90), indicating limited effectiveness under diverse and realistic clinical conditions.

Table 6. Comparison Performance of Segmentation Model on Custom Dataset

Parameter \ Model	YOLOv8n-seg	Hybrid Model	U-NET
Precision	98%	99%	88%
Recall	97%	98%	92%
F1 Score	97%	98%	89%
Dice Score	0.97	0.98	0.9

From the table 6, it has been seen that the hybrid model consistently outperforms both the baseline YOLOv8n-seg model and the U-Net architecture across all evaluation metrics. It achieved the highest precision (99%) and recall (98%), resulting in an F1 score of 98% and a dice score of 0.98 on the custom hand gesture dataset. This highlights the effectiveness of the hybrid model's performance for real-time assistive applications.

The real-time inference results for the hybrid YOLO model after deployment on the embedded device under diverse environmental conditions are shown in Table 7. The table shows the real-time inference results of the hybrid YOLOv8n-seg model integrated with the DETR as the head. This model is deployed on the Raspberry Pi 4B. In the table, the Environment column denotes different real-world testing conditions such as lighting variation, change in hand orientation, motion blur, and gesture overlap with body parts. The Input Frame column shows the corresponding real-time input frames, while the Output Frame column shows the segmented result. From the observations represented in the table, it can be noted that the hybrid model delivers satisfactory performance across these diverse environmental conditions. The framework effectively segments the hand gesture from the other frame details, showing its robustness for real-time deployment.

Table 7. Realtime Inference Result for Hybrid YOLO Model Deployed on Embbeded Device under Diverse Environmental Conditions

Environment	Input Frame	Output Frame	Environment	Input Frame	Output Frame
Light Variation [Low Light]			Light Variation		
Change in Hand Orientation			Gesture Overlap		
Blur			Change in Hand Orientation		

5. Conclusion

This study successfully introduced and validated a hand gesture segmentation framework to address the specific needs of PPD. Due to the absence of publicly available datasets specific to PPD, a custom dataset contacting 11230 images was developed. Using this dataset, the enhanced segmentation capabilities of Transformer-integrated YOLO8n-seg models were effectively demonstrated. The hybrid YOLOv8n-seg model integrated with DETR as a head demonstrated better performance for the segmentation of PPD hand gestures compared to YOLOv8n-seg and U-Net. On the custom dataset, the hybrid model achieved the highest precision (99%), recall (98%), F1 score (98%), and Dice score (0.98), outperforming

both YOLOv8n-seg (98%, 97%, 97%, 97%) and U-Net (88%, 92%, 89%, 0.90) across all evaluation metrics. The model showed satisfactory performance for real time inference when deployed on the Raspberry Pi 4B as an embedded edge device. To our knowledge, this is the first work combining YOLOv8n-seg with a DETR head for PPD hand gesture segmentation. This work provides a robust foundation for the development of practical and deployable hand gesture segmentation and recognition based assistive systems for PPD in different aspects of daily life, including communication and rehabilitation. However, the model response is slightly affected under extreme lighting conditions such as glare or poor illumination, which disturbs the hand boundaries. Thus, further work can focus on improving the system's robustness under extreme light conditions through suitable pre-processing and multi modal fusion.

References

[1] Telepatil, Avadhoot, and Jayashree Vaddin. "Various Optimizers' Performances for CNN-Based Hand Gesture Recognition for PDP Assistance." In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, (2023): 1-7.

[2] Di Libero, Tommaso, Elisa Langiano, Chiara Carissimo, Maria Ferrara, Pierluigi Diotaiuti, and Angelo Rodio. "Technological support for people with Parkinson's disease: A narrative review." Journal of Gerontology and Geriatrics 71 (2023): 87-101.

[3] Deb, Ranadeep, Sizhe An, Ganapati Bhat, Holly Shill, and Umit Y. Ogras. "A systematic survey of research trends in technology usage for Parkinson's disease." Sensors 22, no. 15 (2022): 5491.

[4] Duffy, Orla, Jonathan Synnott, Roisin McNaney, Paola Brito Zambrano, and W. George Kernohan. "Attitudes toward the use of voice-assisted technologies among people with Parkinson disease: findings from a web-based survey." JMIR rehabilitation and assistive technologies 8, no. 1 (2021): e23006.

[5] Kour, Navleen, Sunanda Gupta, and Sakshi Arora. "Sensor technology with gait as a diagnostic tool for assessment of Parkinson's disease: a survey." Multimedia Tools and Applications 82, no. 7 (2023): 10211-10247.

[6] Dineshkumar, V., D. Raveena Judie Dolly, D. J. Jagannath, and J. Dinesh Peter. "Assistive methodologies for Parkinson's disease tremor management—a health opinion." Frontiers in Public Health 10 (2022): 850805.

[7] Deb, Ranadeep, Ganapati Bhat, Sizhe An, Holly Shill, and Umit Y. Ogras. "Trends in technology usage for Parkinson's disease assessment: a systematic review." MedRxiv (2021): 2021-02.

[8] Wang, Peng, Hong Deng, Jiaxu Guo, Siqi Ji, Dan Meng, Jun Bao, and Peng Zuo. "Leaf segmentation using modified YOLOv8-seg models." Life 14, no. 6 (2024): 780.

[9] Alsuwaylimi, Amjad A. "Enhanced YOLOv8-Seg instance segmentation for real-time submerged debris detection." IEEE Access (2024).

[10] Hassanudin, Wahyu Maulana, Victor Gayuh Utomo, and Riski Apriyanto. "Fine-Grained Analysis of Coral Instance Segmentation using YOLOv8 Models." Sinkron: jurnal dan penelitian teknik informatika 8, no. 2 (2024): 1047-1055.

[11] Casas, Edmundo, Leo Ramos, Cristian Romero, and Francklin Rivas-Echeverria. "A comparative study of YOLOv5 and YOLOv8 for corrosion segmentation tasks in metal surfaces." Array 22 (2024): 100351.

[12] Bai, Ruihan, Mingkang Wang, Zhiping Zhang, Jiahui Lu, and Feng Shen. "Automated construction site monitoring based on improved YOLOv8-seg instance segmentation algorithm." IEEE Access 11 (2023): 139082-139096.

[13] Sapkota, Ranjan, Dawood Ahmed, and Manoj Karkee. "Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments." Artificial Intelligence in Agriculture 13 (2024): 84-99.

[14] Li, Xiangtai, Henghui Ding, Haobo Yuan, Wenwei Zhang, Jiangmiao Pang, Guangliang Cheng, Kai Chen, Ziwei Liu, and Chen Change Loy. "Transformer-based visual segmentation: A survey." IEEE transactions on pattern analysis and machine intelligence (2024).

[15] Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. "End-to-end object detection with transformers." In European conference on computer vision, Cham: Springer International Publishing, (2020): 213-229.

[16] Tian, Hao, Yu Cao, and P. Y. Mok. "Detr-based layered clothing segmentation and fine-grained attribute recognition." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2023): 3535-3539.

[17] Zhang, Baocan, Xiaolu Jiang, and Wei Zhao. "An enhanced mask transformer for overlapping cervical cell segmentation based on DETR." IEEE Access (2024).

[18] Jiang, Jiaran, Xin Zuo, Xin Shu, Dan Xu, and Ping Qian. "UW-DETR: Underwater Image Instance Segmentation with Co-DETR." In 2024 7th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), IEEE, (2024): 125-130.

[19] Kadi, Hocine, Théo Sourget, Marzena Kawczynski, Sara Bendjama, Bruno Grollemund, and Agnés Bloch-Zupan. "Segmentation, and Numbering in Oral Rare Diseases: Focus on Data Augmentation and Inpainting Techniques." In 2023 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, (2023): 1358-1363.

[20] Wang, Yucheng, Huiping Wang, Hang Mao, Suwei Gao, Qiaofeng Wei, Shujing Li, Rangtong Liu, and Boyang Xu. "Image Segmentation Method for Water Transport Feature Detection in Fabrics via Target-Located Strategy." IEEE Access (2025).

[21] Tang, Chunming, Xinguang Dai, Jinghong Liu, and Yan Li. "SFED-Former: A Symmetrical Feature Enhanced DETR for Video Lane Detection." In 2023 2nd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR), IEEE, (2023): 121-125.

[22] Bai, Ruihan, Mingkang Wang, Zhiping Zhang, Jiahui Lu, and Feng Shen. "Automated construction site monitoring based on improved YOLOv8-seg instance segmentation algorithm." IEEE Access 11 (2023): 139082-139096.