American Sign Language and Facial Expression Recognition

(Samatha Ereshi Akkamahadevi, Pavan Kumar Reddy Lakkireddy, Ravi Teja Vempati)

Abstract:

In this study, we propose a novel approach for American Sign Language (ASL) and facial expressions detection using deep learning models, specifically YOLOv9 and YOLOv8. The aim of the research is to develop a robust system capable of recognizing ASL gestures and facial expressions in real-time video streams. We evaluated the performance of popular object detection models like : YOLOV9, YOLOV8,DETECTRON2 (FASTER RCNN) and DETR, on a dataset of manually annotated video frames. Our dataset consisted of 4429 training images, 551 validation images, and 195 test images, with a ratio of 22:3:1. We augmented the dataset by applying rotations between 15° and +15° , adding noise up to 3% of pixels, adding brightness between -20% and +20%,and adding exposure between -10% and +10% . We trained, validated, and tested all the models on the dataset and found that YOLOv9 performed better. Our approach involved importing necessary packages, libraries, and configuring the models. We also defined functions for matching detections with tracks based on their IOU overlap. Our experimental setup included reading frames from a video file, running the model on the first frame, formatting the results into detections, annotating the frame with bounding boxes and class labels, and displaying the annotated frame. Our results showed that YOLOv9 was effective in accurately detecting.

Introduction:

The ability to effectively communicate through American Sign Language (ASL) is essential for full societal participation of the deaf and hard-of-hearing community, which makes up 5% of the global population according to the World Health Organization (WHO). With over 466 million people currently affected by disabling hearing loss, and projections indicating this number will surpass 900 million by 2050, there is a pressing need to enhance ASL recognition capabilities. Facial expressions play a crucial role in ASL, featuring a wide range of subtle variations that convey meaning. This project aims to leverage state-of-the-art computer vision models, including YOLOv9, YOLOv8, Detectron, and DETR, to accurately recognize ASL signs and facial expressions in real-time. By improving the recognition of these communicative elements, the project seeks to ensure clear and reliable communication, leading to an improved quality of life and increased opportunities for the deaf and hard-of-hearing community.

Problem Statement:

The objective of this research project is to develop a robust system for real-time American Sign Language (ASL) and facial expression recognition using state-of-the-art deep learning models, including YOLOv9, YOLOv8, Detectron (Faster R-CNN), and DETR (DEtection TRansformer). We aim to address the challenge of accurately detecting and interpreting ASL gestures and facial expressions in dynamic environments, with a focus on improving accessibility and communication for individuals with hearing or speech impairments.

Related Work:

In the realm of American Sign Language (ASL) recognition, significant strides have been made, moving from early glove-based systems to sophisticated computer vision and deep learning techniques. Early research utilized glove-based sensors to capture hand movements, a method noted for its accuracy but criticized for its intrusiveness (Starner, Weaver, & Pentland, 1998). The shift to vision-based systems marked a pivotal development, enabling the non-intrusive capture of ASL signs using cameras, a technique expanded upon by Zafrulla et al. (2011**)** with the implementation of Kinect technology for gesture recognition. Recent advancements have leveraged deep learning to enhance recognition accuracy; Koller et al. (2015) highlighted the use of convolutional and recurrent neural networks to process the spatial and temporal aspects of ASL. The incorporation of object detection algorithms such as Faster R-CNN and YOLO has further refined the real-time performance of ASL recognition systems, offering substantial improvements in speed and accuracy (Ren et al., 2017; Redmon & Farhadi, 2018). Additionally, hybrid models combining CNNs and RNNs have shown great promise in capturing the dynamic nature of sign language (Cui, Liu, & Zhang, 2019). Despite these advancements, challenges persist, particularly in detecting non-manual features like facial expressions and body postures. Ongoing research is directed towards developing multimodal systems that integrate various sensory inputs to enhance recognition capabilities (Jiang et al., 2020). This trajectory from basic sensor-based methods to advanced multimodal deep learning approaches illustrates both the progress and the potential for future breakthroughs in ASL recognition technology.

Experimental Setup:

To conduct the experiments for this research project, the following experimental setup was used:

1. Hardware: The experiments were run on colab with an Intel Xeon CPU @2.20 GHz, 12 GB RAM, and a NVIDIA T4 tensor core gpu specifications.

2. Software: The software used for this project included Python programming language, Google Colab, Roboflow, Ultralytics and supervision package.

3. Dataset: The dataset for the project was manually annotated using Roboflow and consisted of 4429 training images, 551 validation images, and 195 test images.

4. Object Detection Models: Four object detection models, Detectron2 (Faster RCNN), DETR, YOLOv8 and YOLOv9, were used to detect and count vehicles in the video frames.

Technical Approach:

In this research project, the main technical approach was to develop a AI-based solution for Sign Language and Facial Expressions. The project utilized four popular object detection models, namely Detectron2 (Faster RCNN), DETR , YOLOv8 and YOLOv9, to determine the model that performs best in detection. The project's technical approach involved the following steps:

1. Data Collection and Annotation: The first step was to manually annotate the video frames with Roboflow to generate a dataset for training, validation, and testing. The dataset was split into a ratio of 22:3:1 for training, validation, and testing. Furthermore, the dataset was augmented by applying rotations between 15° and +15° , adding noise up to 3% of pixels, adding brightness between -20% and +20%,and adding exposure between -10% and +10%.
2. Model Training: The next step involved training the four object detection models using the annotated dataset. Detectron2 (Faster RCNN), DETR, YOLOv8 and YOLOv9 models were trained with specific parameters, including IMS_PER_BATCH = 4, BASE_LR = 0.0025, WARMUP_ITERS = 500, MAX_ITER = 12000, STEPS = (1000,3000,6000) for Detectron2 (Faster RCNN) and LR = 1e-4, LR_BACKBONE = 1e-5, WEIGHT_DECAY = 1e-4, IOU_THRESHOLD = 0.5 for DETR and Epochs = 25, Batch_size = 16, Imgsz = 640 pixels, for YOLOv8 and Epochs = 20, Batch_size = 16, Imgsz = 640 pixels for YOLOv9. The performance of all these models was evaluated to determine the model that performs best.
3. Model Selection: From the results obtained during model training and validation, it was concluded that the YOLOv9 model performed better in detecting.

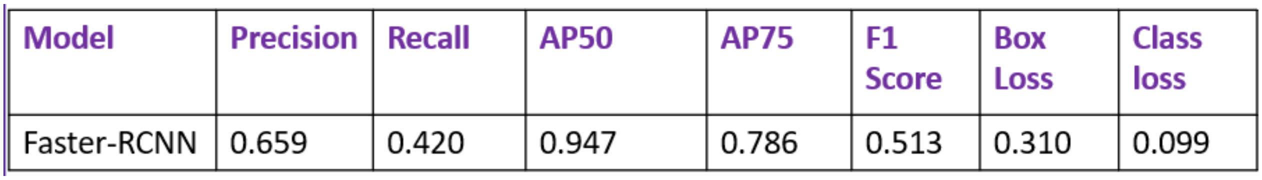
Results:

The results obtained from the experiments conducted for this research project showed that the YOLOv9 model outperformed the Detectron2 (Faster RCNN), DETR and YOLOv8 model in detecting signd and facial expressions in the video frames. The YOLOv9 model was able to detect signs and expressions accurately and efficiently, with a higher accuracy rate than the Detectron2 (Faster RCNN), DETR, YOLOv8 model

On training and testing all these detectron 2 (Faster RCNN), DETR, YOLOv8 and YOLOv9 on our custom data set, we obtained the following results

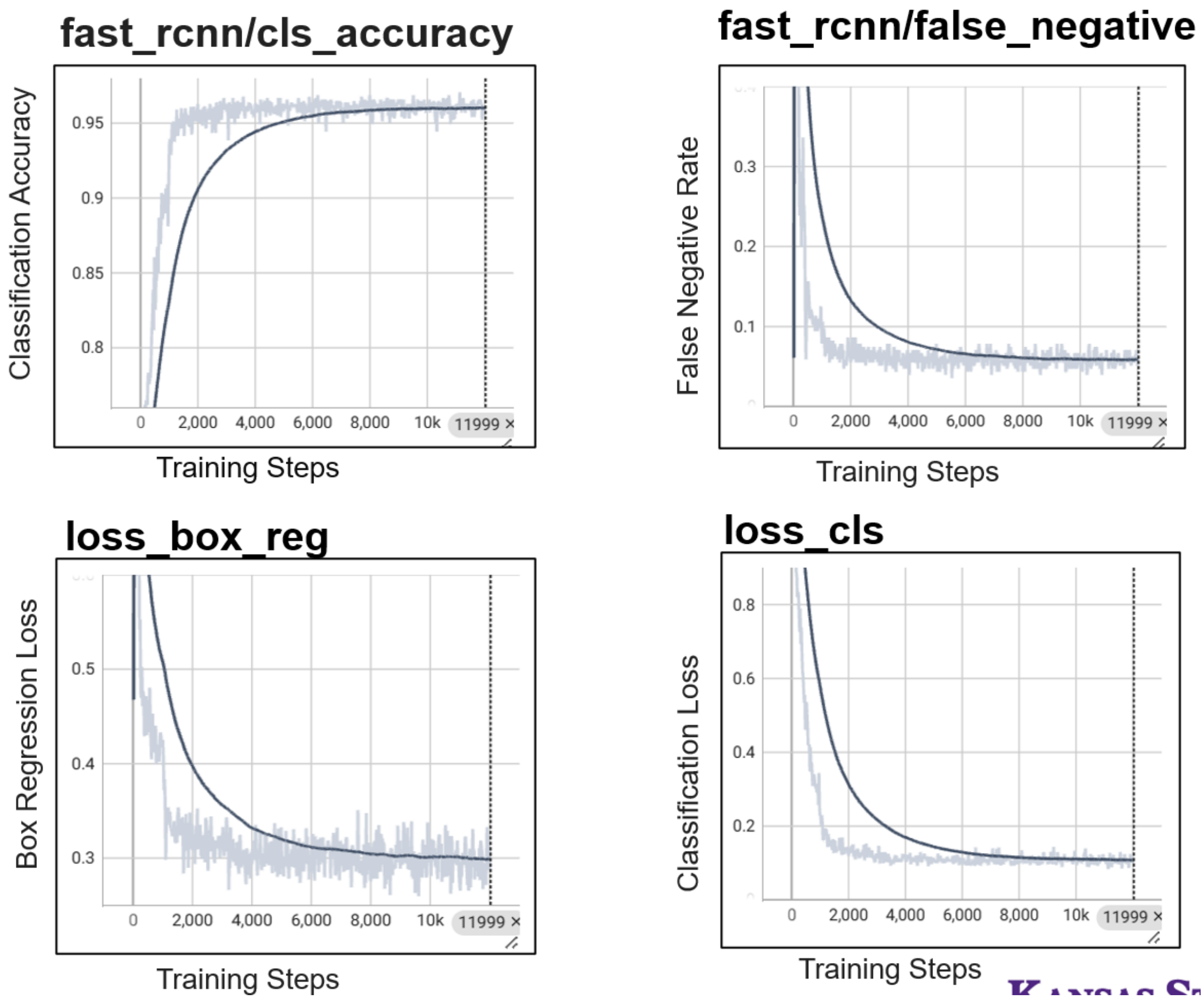
Training time of Detectron2 (Faster RCNN) : 40 minutes.

Test results of Detectron2 (Faster RCNN)



Model	Precision	Recall	AP50	AP75	F1 Score	Box Loss	Class loss
Faster-RCNN	0.659	0.420	0.947	0.786	0.513	0.310	0.099

Graphs of Detectron2 (Faster RCNN)

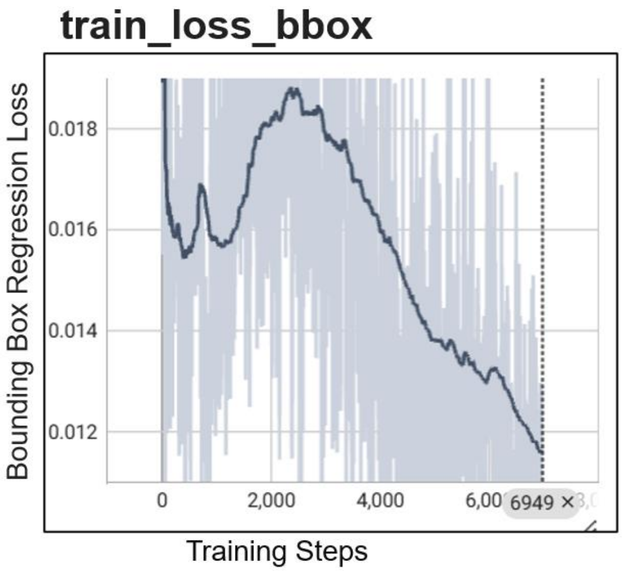


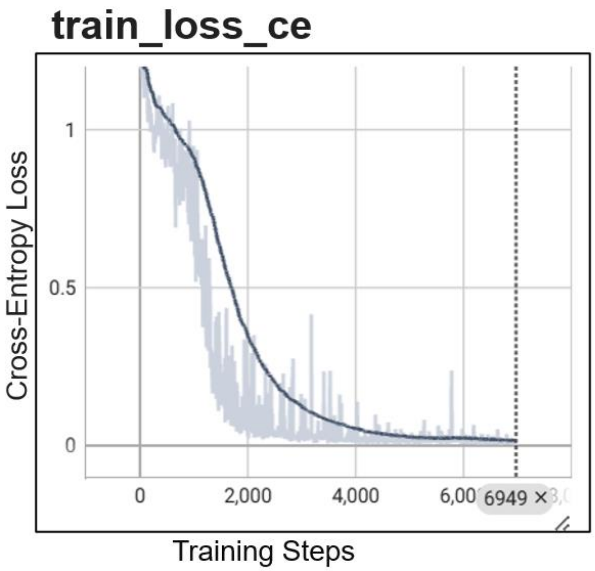
Training time of Detectron2 (Faster RCNN) : 4hrs.

Test results of DETR

Model	Precision	Recall	AP50	AP75	F1 Score	Box Loss	Cross Entropy loss
DETR	0.847	0.358	0.545	0.656	0.504	0.009	0.018

Graphs of DETR



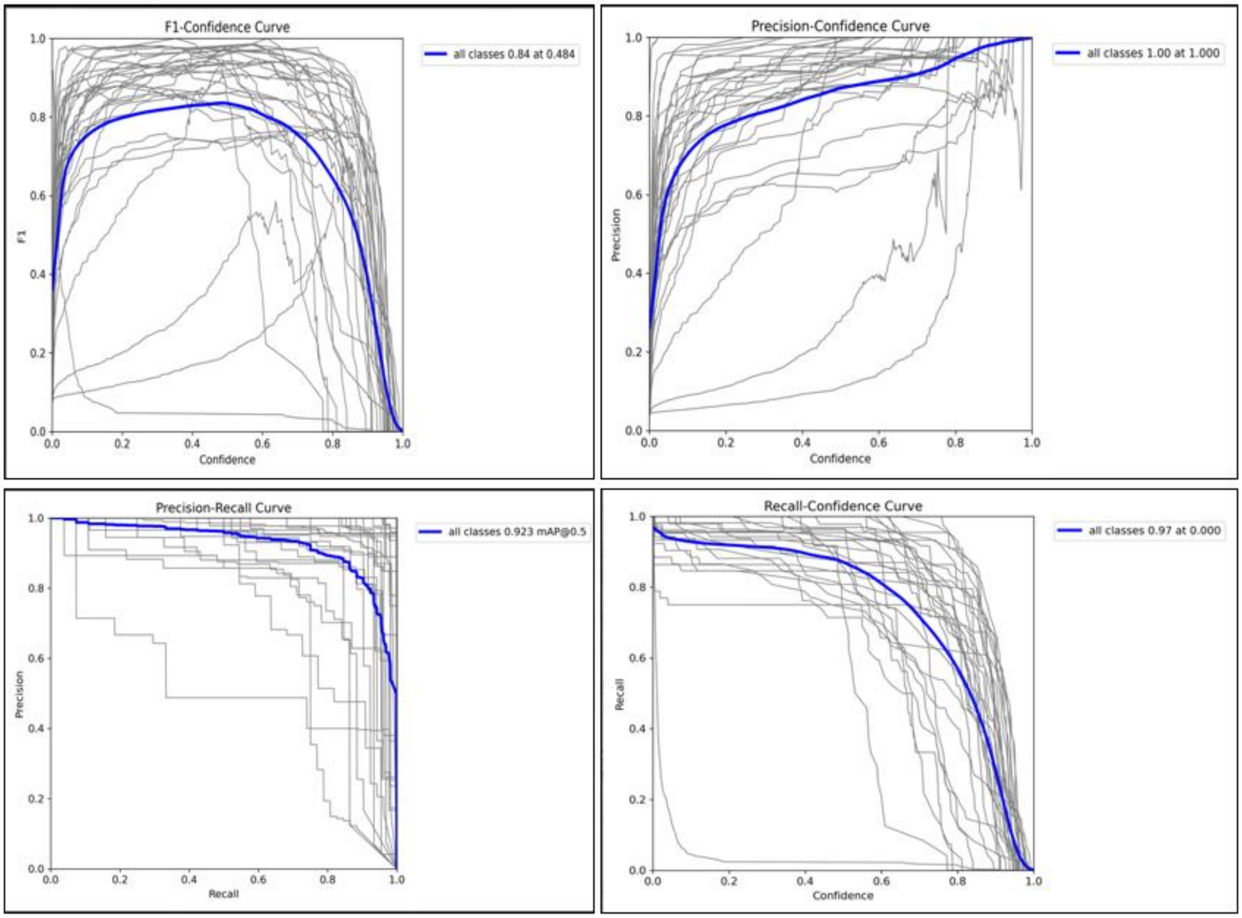


Training time of YOLOv8 : 1 hr 45 min.

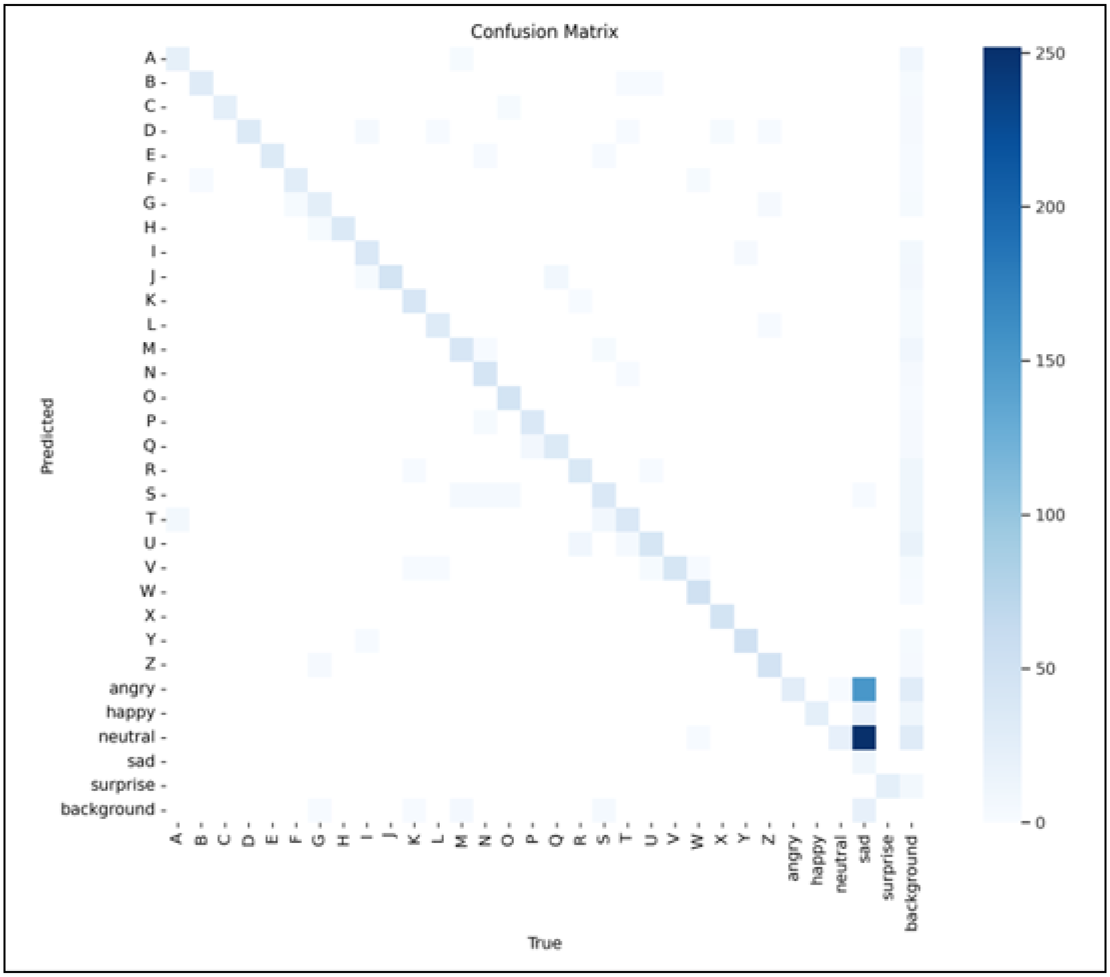
Test results of YOLOV8

Model	Precision	Recall	mAP50	mAP 50-95	F1 Score	Box Loss	Class Loss	DFL Loss
YOLOv8	0.863	0.882	0.923	0.602	0.872	0.94	0.478	0.872

YOLOV8 test result graphs



Confusion Matrix

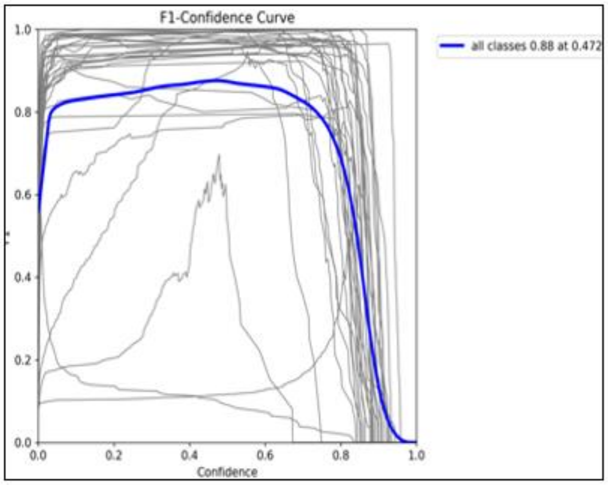


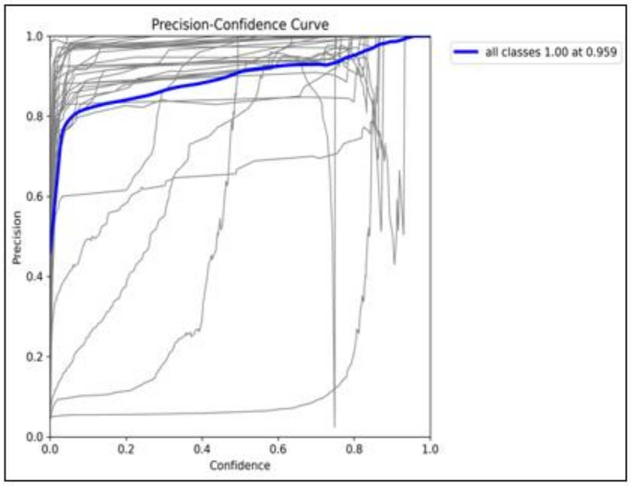
Training time of YOLOv9 : 2hrs.

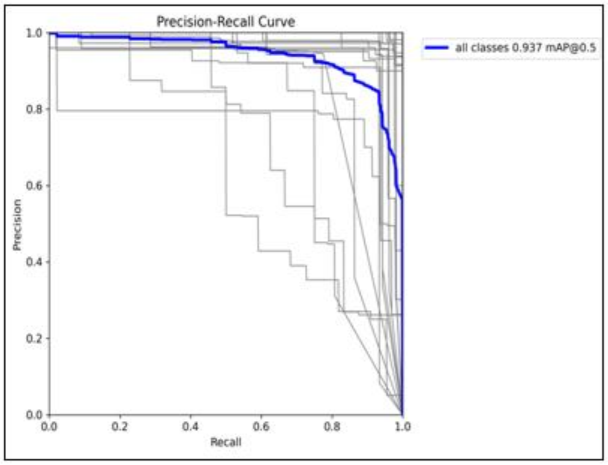
Test results of YOLOv9

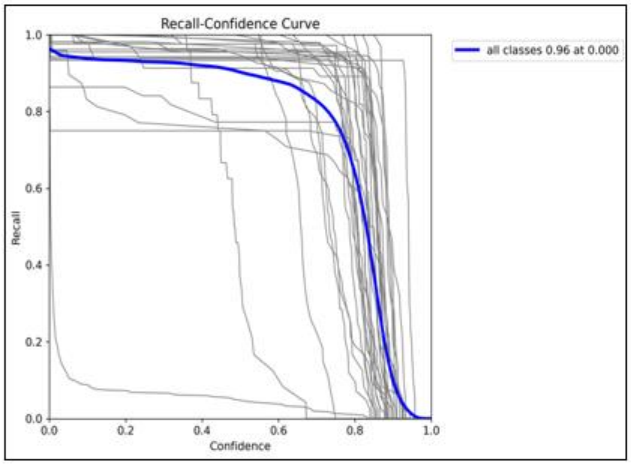
Model	**Precision**	**Recall**	**mAP50**	**mAP 50-95**	**F1 Score**	**Box Loss**	**Class Loss**	**DFL Loss**
YOLOv9	0.896	0.913	0.937	0.620	0.904	1.024	0.462	1.21

YOLOV9 test result graphs

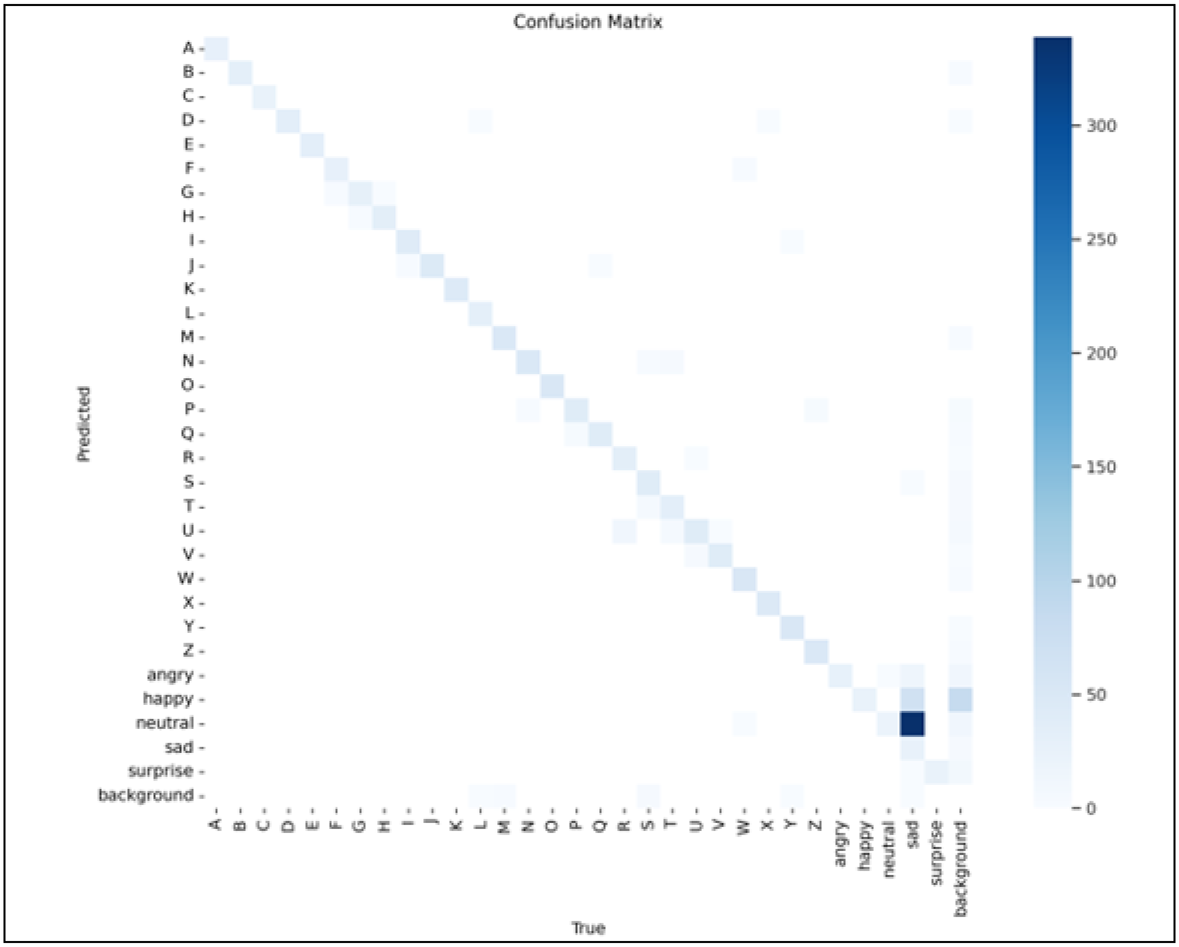




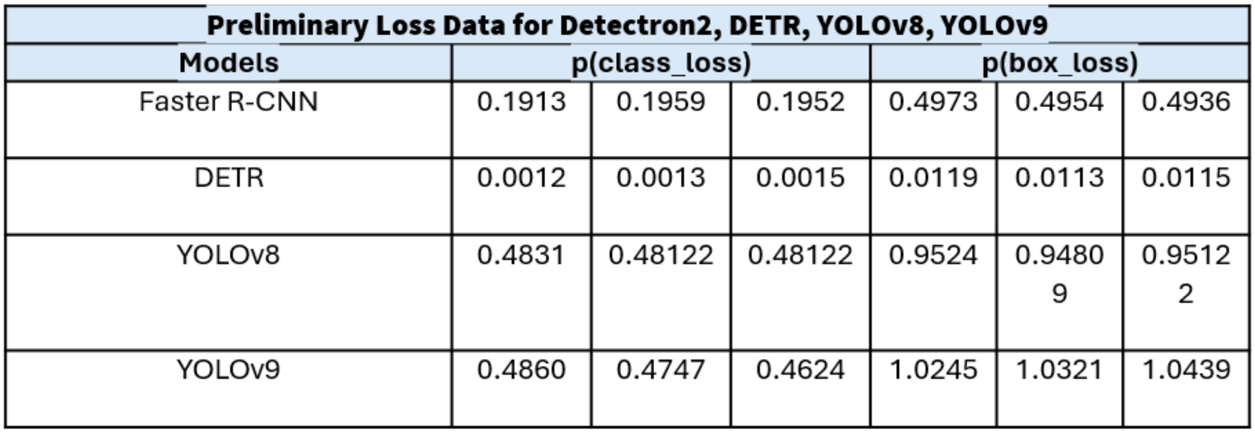




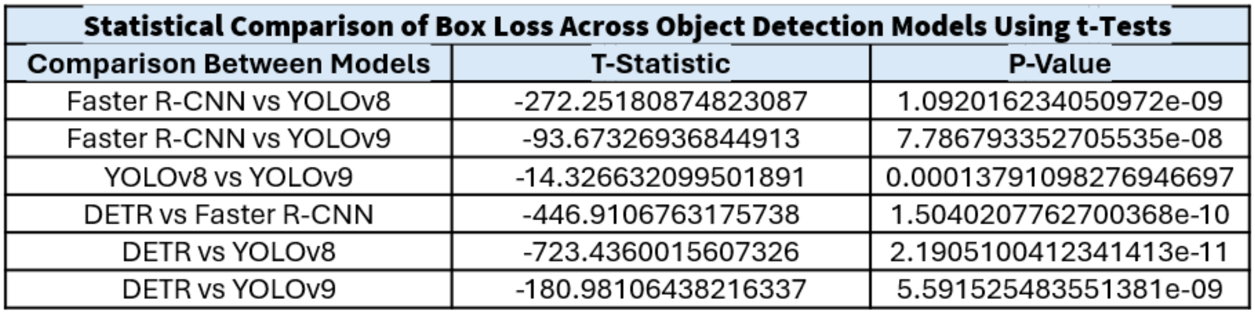
Confusion Matrix



Hypothesis Testing:



Preliminary Loss Data for Detectron2, DETR, YOLOv8, YOLOv9						
Models	**p(class_loss)**			**p(box_loss)**		
Faster R-CNN	0.1913	0.1959	0.1952	0.4973	0.4954	0.4936
DETR	0.0012	0.0013	0.0015	0.0119	0.0113	0.0115
YOLOv8	0.4831	0.48122	0.48122	0.9524	0.94809	0.95122
YOLOv9	0.4860	0.4747	0.4624	1.0245	1.0321	1.0439



Statistical Comparison of Box Loss Across Object Detection Models Using t-Tests		
Comparison Between Models	**T-Statistic**	**P-Value**
Faster R-CNN vs YOLOv8	-272.25180874823087	1.092016234050972e-09
Faster R-CNN vs YOLOv9	-93.67326936844913	7.786793352705535e-08
YOLOv8 vs YOLOv9	-14.326632099501891	0.00013791098276946697
DETR vs Faster R-CNN	-446.9106763175738	1.5040207762700368e-10
DETR vs YOLOv8	-723.4360015607326	2.1905100412341413e-11
DETR vs YOLOv9	-180.98106438216337	5.591525483551381e-09

1. **Faster R-CNN vs YOLOv8:**
   - **Interpretation:** The negative t-statistic (-272.25) indicates that the mean box loss score of Faster R-CNN is significantly lower than that of YOLOv8. The extremely small p-value (1.09e-09) confirms a statistically significant difference. This suggests that Faster R-CNN performs better than YOLOv8 in terms of box prediction accuracy.
2. **Faster R-CNN vs YOLOv9:**
   - **Interpretation:** Similarly, the negative t-statistic (-93.67) indicates that Faster R-CNN achieves a significantly lower mean box loss score compared to YOLOv9. The very small p-value (7.79e-08) confirms a statistically significant difference, implying that Faster R-CNN outperforms YOLOv9 in terms of box prediction accuracy.
3. **YOLOv8 vs YOLOv9:**
   - **Interpretation:** The negative t-statistic (-14.33) indicates that the mean box loss score for YOLOv8 is significantly lower than that for YOLOv9. Although the p-value (0.00014) is relatively small, it still confirms a statistically significant difference, suggesting that YOLOv8 performs better than YOLOv9 in terms of box prediction accuracy.
4. **DETR vs Faster R-CNN:**
   - **Interpretation:** The negative t-statistic (-446.91) indicates that the mean box loss score of DETR is significantly lower than that of Faster R-CNN. The extremely small p-value (1.50e-

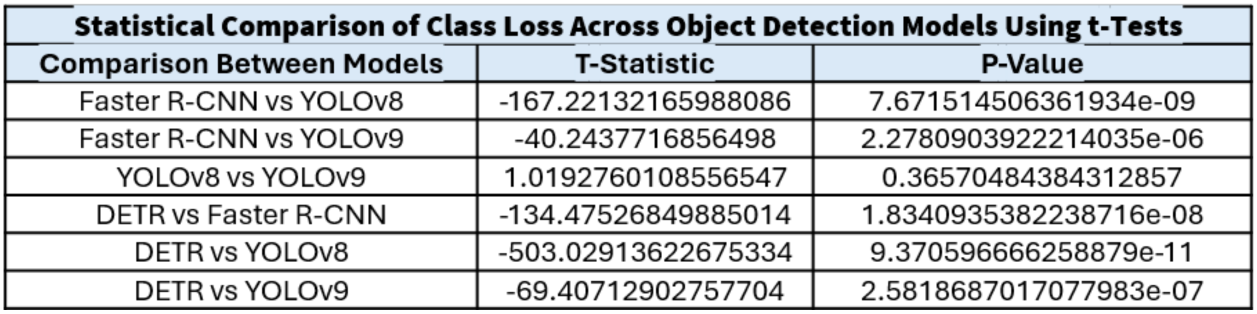
10) confirms a statistically significant difference, implying that DETR performs better than Faster R-CNN in terms of box prediction accuracy.

5. **DETR vs YOLOv8:**

    - **Interpretation:** Similarly, the negative t-statistic (-723.44) indicates that DETR achieves a significantly lower mean box loss score compared to YOLOv8. The extremely small p-value (2.19e-11) confirms a statistically significant difference, suggesting that DETR outperforms YOLOv8 in terms of box prediction accuracy.

6. **DETR vs YOLOv9:**

    - **Interpretation:** The negative t-statistic (-180.98) indicates that DETR achieves a significantly lower mean box loss score compared to YOLOv9. The very small p-value (5.59e-09) confirms a statistically significant difference, implying that DETR performs better than YOLOv9 in terms of box prediction accuracy.



Statistical Comparison of Class Loss Across Object Detection Models Using t-Tests		
Comparison Between Models	**T-Statistic**	**P-Value**
Faster R-CNN vs YOLOv8	-167.22132165988086	7.671514506361934e-09
Faster R-CNN vs YOLOv9	-40.2437716856498	2.2780903922214035e-06
YOLOv8 vs YOLOv9	1.0192760108556547	0.36570484384312857
DETR vs Faster R-CNN	-134.47526849885014	1.8340935382238716e-08
DETR vs YOLOv8	-503.02913622675334	9.370596666258879e-11
DETR vs YOLOv9	-69.40712902757704	2.5818687017077983e-07

1. **Faster R-CNN vs YOLOv8:**

    - **Interpretation:** The negative t-statistic (-167.22) indicates that the mean class loss score of Faster R-CNN is significantly lower than that of YOLOv8. The extremely small p-value (7.67e-09) confirms a statistically significant difference, suggesting that Faster R-CNN performs better than YOLOv8 in terms of class prediction accuracy.

2. **Faster R-CNN vs YOLOv9:**

    - **Interpretation:** Similarly, the negative t-statistic (-40.24) indicates that Faster R-CNN achieves a significantly lower mean class loss score compared to YOLOv9. The very small p-value (2.28e-06) confirms a statistically significant difference, implying that Faster R-CNN outperforms YOLOv9 in terms of class prediction accuracy.

3. **YOLOv8 vs YOLOv9:**

    - **Interpretation:** The positive t-statistic (1.02) indicates that there is no significant difference in the mean class loss scores between YOLOv8 and YOLOv9. With a p-value of 0.37, which is greater than 0.05, this difference is not statistically significant. Therefore, YOLOv8 and YOLOv9 perform similarly in terms of class prediction accuracy.

4. **DETR vs Faster R-CNN:**

   - **Interpretation:** The negative t-statistic (-134.48) indicates that the mean class loss score of DETR is significantly lower than that of Faster R-CNN. The extremely small p-value (1.83e-08) confirms a statistically significant difference, implying that DETR performs better than Faster R-CNN in terms of class prediction accuracy.

5. **DETR vs YOLOv8:**

   - **Interpretation:** Similarly, the negative t-statistic (-503.03) indicates that DETR achieves a significantly lower mean class loss score compared to YOLOv8. The extremely small p-value (9.37e-11) confirms a statistically significant difference, suggesting that DETR outperforms YOLOv8 in terms of class prediction accuracy.

6. **DETR vs YOLOv9:**

   - **Interpretation:** The negative t-statistic (-69.41) indicates that DETR achieves a significantly lower mean class loss score compared to YOLOv9. The very small p-value (2.58e-07) confirms a statistically significant difference, implying that DETR performs better than YOLOv9 in terms of class prediction accuracy.

Conclusion:

In conclusion, this research project demonstrated the effectiveness of using AI-based solutions for sign detection and facial expression. The project utilized four popular object detection models, Detectron2 (Faster RCNN), DETR, YOLOv8 and YOLOv9, to determine the model that performs best in detection . The results obtained from the experiments showed that YOLOv9 performed better for detection than detectron 2 **(**AP50 of Detectron2 (Faster Rcnn) : 65.9) (AP50 of DETR : 84.7) (AP50 of YOLOV8 : 92.3)**.**

References:

- Starner, T., Weaver, J., & Pentland, A. (1998). Real-time American sign language recognition using desk and wearable computer-based video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12), 1371-1375.
- Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., & Presti, P. (2011). American Sign Language recognition with the Kinect. Proceedings of the 13th international conference on multimodal interfaces, 279-286.
- Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2015). Deep learning of mouth shapes for sign language. Proceedings of the 3rd Workshop on Recognizing Detailed Human Actions at CVPR.

- Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137-1149.
- Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.
- Cui, R., Liu, H., & Zhang, C. (2019). A deep neural network for real-time detection and classification of sign language. IEEE Access, 7, 116753-116762.
- Jiang, F., Zhang, S., Wu, S., Gao, Y., & Zhao, D. (2020). Deep multimodal learning for real-time detection of non-manual features in sign language. IEEE Transactions on Multimedia, 22(3), 707-720.