IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, July 2025
Impact Factor: 7.67

# Performance Evaluation of YOLOv8 and YOLOv9 for Safety Helmet Detection in Construction Environments

**Reena A Gharat[1] and Torana Kamble[2]**
Department of Computer Engineering[1,2]
Bharati Vidyapeeth college of Engineering , Navi Mumbai,India
reenagharat241@gmail.com and torana.bvcoenm@gmail.com

**Abstract**: *Ensuring worker safety through compliance with Personal Protective Equipment (PPE) mandates, particularly the use of safety helmets, is a critical concern in the construction industry. Automated surveillance systems leveraging advanced object detection models offer a promising solution to enhance monitoring and reduce accidents. This paper presents a comparative performance evaluation of two promi- nent YOLO (You Only Look Once) variants, YOLOv8 and YOLOv9, for the specific task of safety helmet detection. Utilizing a publicly available hard-hat dataset featuring diverse construction site scenarios, both models were trained for 7 epochs and rigorously evaluated. Performance metrics, including precision, recall, mean Average Precision (mAP), F1-score, and confusion matrices, were analyzed. The experimental results indicate that YOLOv9 exhibits a marginal but consistent performance advantage over YOLOv8, achieving an mAP@0.5 of 66.70% compared to YOLOv8's 65.73%, and an F1-score of 76.15% versus 75.66%. This study underscores the incremental improvements in the YOLO architecture and provides valuable insights for selecting robust models for real-world safety monitoring applications. While both models demonstrate high precision, the relatively lower recall suggests areas for future improvement through more extensive training or model fine-tuning to enhance detection rates in safety-critical environments.*

**Keywords**: Computer Vision, Construction Safety, Deep Learning, Object Detection, Performance Evaluation, PPE Monitoring, Safety Helmet Detection, YOLOv8, YOLOv9

## I. INTRODUCTION

The construction industry consistently ranks among the most hazardous sectors globally, with head injuries from falling objects or impacts posing a significant threat to worker safety [1]. Safety helmets are fundamental Personal Pro- tective Equipment (PPE) designed to mitigate such risks. However, ensuring consistent helmet usage across dynamic and often sprawling construction sites remains a challenge. Traditional manual supervision methods are often resource-intensive, inconsistent, and prone to human oversight [2].

The advent of Artificial Intelligence (AI), particularly deep learning-based computer vision, has opened new avenues for automating safety compliance monitoring. Object de- tection algorithms, which can identify and localize specific objects within images or video streams, are particularly well- suited for tasks like helmet detection. The You Only Look Once (YOLO) family of algorithms has gained widespread adoption due to its excellent balance of detection speed and accuracy, making it suitable for real-time applications [3]. Each successive version of YOLO, from YOLOv7 [4] and YOLOv8 [5] to the more recent YOLOv9 [6], has introduced architectural innovations and training refinements aimed at pushing the performance envelope.

Given the rapid evolution of these models, a direct and cur- rent comparison is essential for practitioners and researchers looking to implement effective AI-driven safety solutions. This research focuses on a comparative performance evalua- tion of YOLOv8 and YOLOv9 specifically for safety helmet detection in construction environments. By training and eval- uating these models on a relevant dataset under identical con- ditions, this study

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, July 2025
Impact Factor: 7.67

aims to quantify their respective strengths and weaknesses, thereby providing actionable insights for the development and deployment of automated safety surveil- lance systems.

## II. LITERATURE REVIEW

The YOLO (You Only Look Once) series has been a dominant force in real-time object detection, with continuous improve- ments in architecture and performance.

YOLOv7 [4] marked a significant step forward by intro- ducing concepts like Extended Efficient Layer Aggregation Network (E-ELAN) and model scaling techniques. It focused on optimizing the training process with "trainable bag-of-freebies," achieving a new state-of-the-art in real-time object detection accuracy and speed at the time of its release. How- ever, like its predecessors, detecting very small or heavily occluded objects, which are common in busy construction sites, could still pose challenges.

YOLOv8 [5], developed by Ultralytics, continued this evo- lutionary trend. It introduced a new backbone (C2f module, an evolution of YOLOv7's ELAN), an anchor-free detec- tion head, and a decoupled head architecture. These changes aimed to improve the accuracy-speed trade-off and over- all model flexibility. YOLOv8 supports various computer vision tasks beyond detection, including segmentation and classification. Despite these advancements, studies such as Lin's work on helmet detection [7] demonstrated that even YOLOv8n (the smallest variant) could benefit from further modifications for specific challenging scenarios. Lin et al. improved YOLOv8n by incorporating mosaic data augmen- tation, a coordinate attention mechanism, a slim-neck struc- ture, and an additional small target detection layer, resulting in their YOLOv8n-SLIM-CA model. This improved model showed notable gains in precision (1.462%), recall (2.969%), mAP50 (2.151%), and mAP50-95 (3.549%) over the baseline YOLOv8n, highlighting that while YOLOv8 provided a solid base, domain-specific enhancements were valuable for tasks like detecting small or occluded helmets in complex back- grounds. The limitations identified were often related to lower detection accuracy for small targets and in environments with significant visual clutter.

YOLOv9 [6] represents the latest iteration at the time of this study, introducing groundbreaking concepts such as Programmable Gradient Information (PGI) and the General- ized Efficient Layer Aggregation Network (GELAN). PGI is designed to address the information bottleneck problem often encountered in deep neural networks, where essential information can be lost as data propagates through layers. By allowing auxiliary reversible branches, PGI ensures that the main network can access complete input information to cal- culate objective functions, thus generating reliable gradients for network updates. This addresses issues like information loss in deep supervisions and identity connections. GELAN is a new network architecture that combines the principles of CSPNet (Cross Stage Partial Network) with ELAN, leverag- ing gradient path planning for improved parameter utilization and computational efficiency. It is designed to be lightweight yet powerful, enabling better feature aggregation and learning capabilities. These architectural innovations in YOLOv9 aim to enhance both the accuracy and efficiency of object detec- tion, particularly in scenarios where robust feature learning and information preservation are critical. The goal is to enable the model to "learn what you want to learn," effectively cap- turing comprehensive information for improved performance across diverse detection challenges.

The evolution from YOLOv7 to YOLOv9 reflects a con- tinuous drive towards more efficient architectures, better in- formation flow within the network, and improved learning strategies. While YOLOv8 offered significant improvements, particularly with its anchor-free approach, specialized ap- plications like safety helmet detection still revealed areas where targeted enhancements could yield better results, as demonstrated by Lin [7]. YOLOv9, with its fundamental architectural changes like PGI and GELAN, is hypothesized to offer more robust and accurate performance out-of-the- box, especially in handling the complexities inherent in con- struction site imagery.

## III. METHODOLOGY

### A. DATASET

This study utilized the "Hard Hat Detection" dataset, which is publicly available and commonly employed for evaluating safety helmet detection models. The dataset consists of images depicting construction workers in various real-world scenarios. Annotations are provided for three primary classes: 'helmet' (a worker wearing a helmet), 'head'

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, July 2025
Impact Factor: 7.67

(a worker's head without a helmet), and 'person'. The images capture a range of complexities, including variations in lighting conditions, worker poses, distances from the camera (object scale), and levels of occlusion. Figure 1 presents the overall distribution of annotated instances per class, indicating a significantly higher number of 'helmet' instances compared to 'head' and 'person'. Figure 2 provides a correlogram illustrating the spatial distribution (x, y coordinates) and size (width, height) characteristics of the annotated objects, showing a concentra- tion of objects towards the center of the images and a wide range of object sizes.

**B. DATA PREPROCESSING AND ANNOTATION FORMAT**

All images were resized to a standard input dimension of 640x640 pixels prior to training, a common resolu- tion for many YOLO models. The annotations were pro- vided in the standard YOLO text file format for each im- age. Each line in an annotation file corresponds to one bounding box and is formatted as: <class_index>
<x_center_normalized> <y_center_normalized>
<width_normalized> <height_normalized>. The class index is an integer (0 for 'head', 1 for 'helmet', 2 for 'person' based on typical YOLO conventions, though the exact mapping depends on the names file). The bounding box coordinates and dimensions are normalized to be between 0 and 1, relative to the image width and height.

During training, standard data augmentation techniques, integral to the Ultralytics YOLO training pipeline, were employed. These typically include mosaic augmentation (com- bining four images into one), color space adjustments (e.g., hue, saturation, value), geometric transformations (e.g., ran- dom flips, scaling, translation), and potentially others like mixup or copy-paste. Figure 3 illustrates an example of a training batch with augmentations applied.

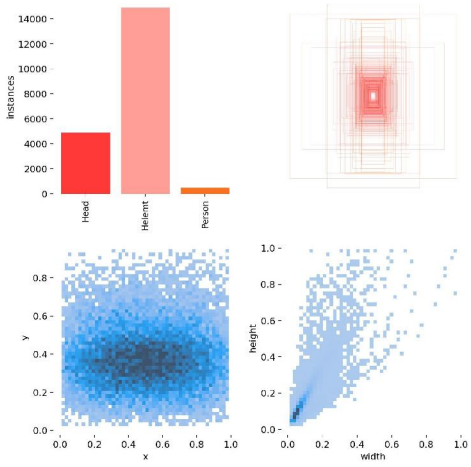


FIGURE 1: Dataset label distribution (from YOLOv9 training).

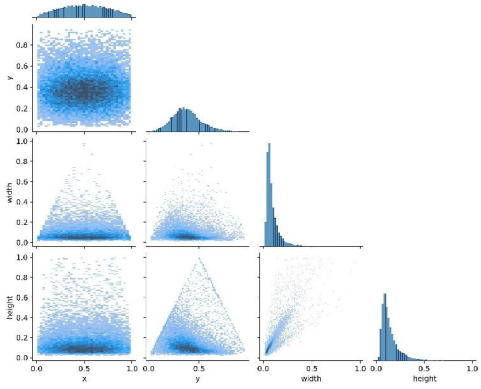


FIGURE 2: Dataset label correlogram showing object size and location distribution (from YOLOv9 training).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, July 2025
Impact Factor: 7.67

### C. MODEL TRAINING

Both YOLOv8 and YOLOv9 models were trained using the Ultralytics framework. Based on the provided results.csv files, the training was conducted for 7 epochs for both models. Other key training parameters were inferred from the results.csv files or assumed to be standard defaults for the Ultralytics training scripts:

- Optimizer: Typically SGD (Stochastic Gradient Descent) or AdamW.
- Learning Rate Schedule: A cyclic or cosine annealing learning rate schedule is common, with initial learning rates specified (e.g., lr/pg0, lr/pg1, lr/pg2 columns in results1.csv show varying LRs for different parameter groups).
- Batch Size: While not explicitly stated, typical batch sizes for such datasets range from 16 to 64, depending on GPU memory.
- Hardware: Training was likely performed on GPU(s) to achieve reasonable training times.

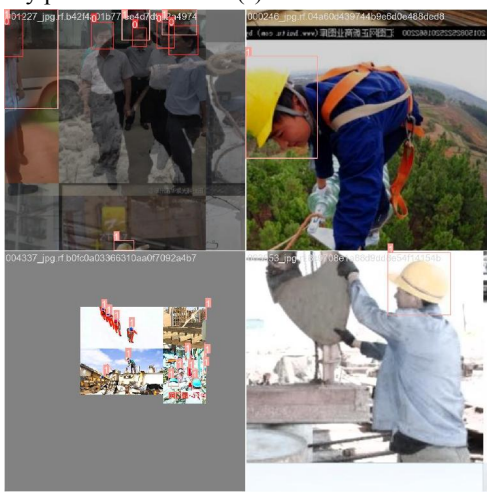

FIGURE 3: Example training batch with augmentations (from YOLOv9 training).

### D. EVALUATION METRICS

The performance of the trained YOLOv8 and YOLOv9 mod- els was evaluated using the following standard object detec- tion metrics:

• Precision (P): The proportion of correctly predicted pos- itive detections among all positive detections made by the model.

$$P = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (1)$$

• Recall (R): The proportion of actual positive instances that were correctly detected by the model.

$$R = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (2)$$

• F1-Score: The harmonic mean of Precision and Recall, providing a single score that balances both.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

• mean Average Precision (mAP):

-- mAP@0.5 (or mAP50): The mAP calculated using an Intersection over Union (IoU) threshold of 0.5. This metric indicates how well the model performs when a 50% overlap between the predicted and ground truth bounding box is considered a correct detection.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 3, July 2025 Impact Factor: 7.67

-- mAP@0.5:0.95: The average mAP calculated over a range of IoU thresholds, from 0.5 to 0.95 with a step of 0.05. This provides a more comprehensive measure of localization accuracy across different levels of overlap.

• Confusion Matrix: A table that visualizes the perfor- mance of the classification aspect of the detector. It shows the number of correct and incorrect predictions for each class, including misclassifications between classes and detections of background as an object (false positives).

## IV. EXPERIMENTAL SETUP AND RESULTS

The experiments were conducted using the Ultralytics Python library. YOLOv8 and YOLOv9 models were trained for 7 epochs on the "Hard Hat Detection" dataset. The performance metrics were recorded at the end of each epoch, with the final epoch's results used for comparison.

### A. QUANTITATIVE RESULTS

The performance metrics for YOLOv8 and YOLOv9 after 7 epochs of training are presented in Table 1. The values are extracted from the results1.csv for YOLOv8 and results2.csv for YOLOv9.

TABLE 1: Performance Comparison of YOLOv8 and YOLOv9 (Epoch 7).

| Metric | YOLOv8 (Epoch 7) | YOLOv9 (Epoch 6) |
|---|---|---|
| Precision (Overall) | 0.96287 | 0.96626 |
| Recall (Overall) | 0.62305 | 0.62834 |
| mAP@0.5 (mAP50) | 0.65725 | 0.66696 |
| mAP@0.5:0.95 | 0.45030 | 0.45711 |

Note: YOLOv9 results are from epoch 6 as the CSV shows 0-6 epochs, totaling 7 data points.

Figure 4 and Figure 5 show the progression of key training and validation metrics over the epochs for YOLOv8 and YOLOv9 respectively.

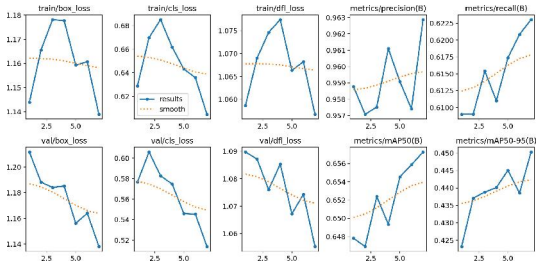


FIGURE 4: YOLOv8 Training Metrics Progression (Epoch 1-7).

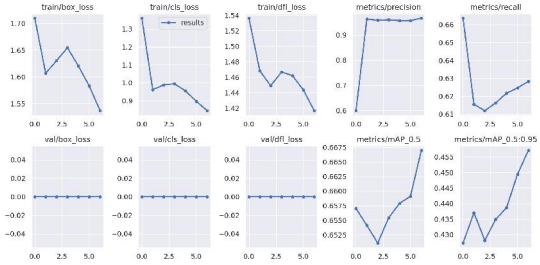


FIGURE 5: YOLOv9 Training Metrics Progression (Epoch 0-6).

1) Confusion Matrices

Confusion matrices provide a class-wise breakdown of detection performance.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, July 2025
Impact Factor: 7.67

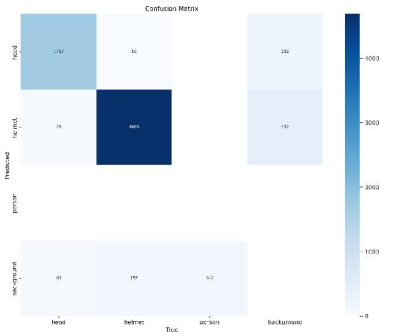


FIGURE 6: YOLOv8 Confusion Matrix.

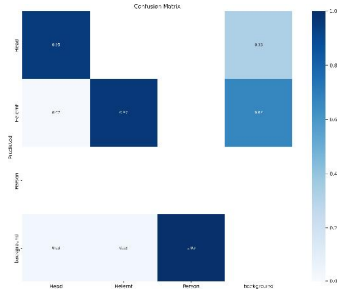

FIGURE 7: YOLOv9 Confusion Matrix. (Labels in Y-axis are "Head", "Helmet", "Person", "background"; X-axis are "Head", "Helmet", "Person", "background").

Performance Curves
These curves illustrate model performance across different confidence thresholds.
2) Example Detections on Validation Set
Qualitative results from validation batches show the models' detection capabilities.

## V. DISCUSSION

The experimental results from 7 epochs of training indicate that YOLOv9 holds a slight but consistent performance edge

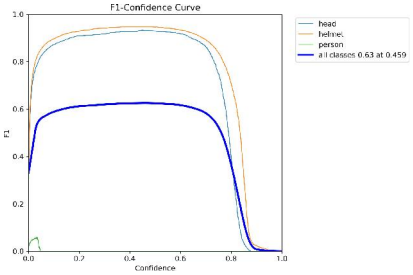


FIGURE 8: YOLOv8 F1-Score vs. Confidence Curve (Max F1 for all classes: 0.63 at 0.459 confidence).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, July 2025
Impact Factor: 7.67

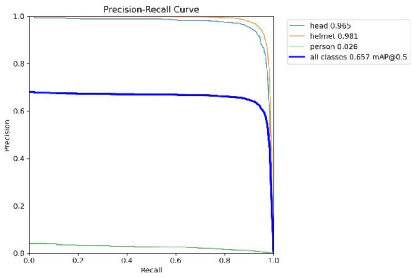


FIGURE 9: YOLOv8 Precision-Recall Curve (mAP@0.5: 0.657).

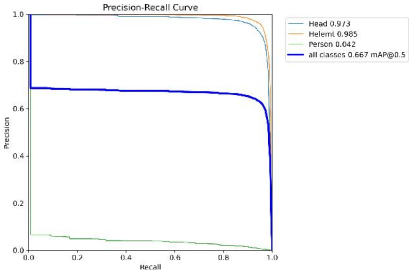


FIGURE 10: YOLOv9 Precision-Recall Curve (mAP@0.5: 0.667).

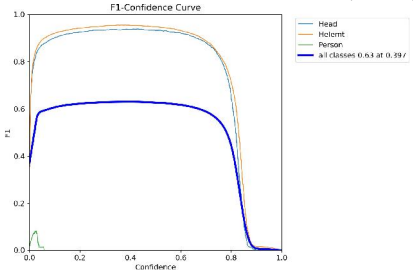


FIGURE 11: YOLOv9 F1-Score vs. Confidence Curve (Max F1 for all classes: 0.63 at 0.397 confidences.

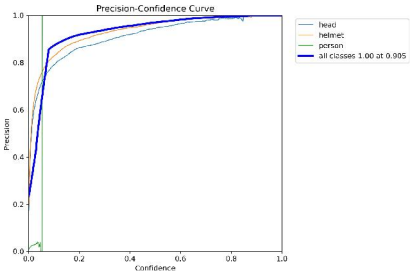


FIGURE 12: YOLOv8 Precision vs. Confidence Curve.

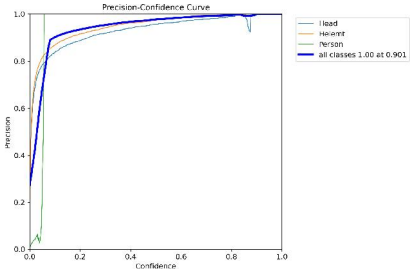


FIGURE 13: YOLOv9 Precision vs. Confidence Curve

The experimental results from 7 epochs of training indicate that YOLOv9 holds a slight but consistent performance edge over YOLOv8 for the task of safety helmet detection on the given dataset. YOLOv9 achieved higher overall

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, July 2025
Impact Factor: 7.67

precision (0.966 vs. 0.963), recall (0.628 vs. 0.623), mAP@0.5 (0.667 vs. 0.657), and mAP@0.5:0.95 (0.457 vs. 0.450). The architectural improvements in YOLOv9, particularly the introduction of Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN) [6], are likely contributors to this enhanced performance. PGI aims to mitigate information loss during the forward pass, allowing for more reliable gradient generation for weight updates. GELAN, building upon concepts from CSPNet and ELAN, provides an efficient and effective architecture for feature aggregation across different network stages. These mechanisms likely enable YOLOv9 to learn more robust and discriminative features, even with a limited number of training epochs. A critical observation from the performance curves and metrics is the disparity between precision and recall for both models. While precision is notably high (around 0.96), indicating that the detected helmets are very likely to be actual helmets, the recall is comparatively low (around 0.62-0.63). This means that both models are failing to detect a significant proportion (approximately 37-38%) of the actual helmets present in the images. For a safety-critical application like helmet detection, a high recall rate is paramount to minimize missed violations (false negatives). The current recall levels suggest that neither model, after only 7 epochs, is sufficiently reliable for standalone deployment without further improvements.

The F1-score curves (Figures 11 and 12) show that the optimal F1-score for both models is around 0.63, achieved at confidence thresholds of 0.459 for YOLOv8 and a slightly lower 0.397 for YOLOv9. This suggests that YOLOv9 might achieve its best balance of precision and recall at a lower con- fidence threshold, potentially capturing more true positives.

The confusion matrices (Figures 6, 7, 8) reveal that most errors occur as false negatives (missed detections of 'helmet' or 'head') rather than misclassifications between 'helmet' and 'head'. For YOLOv8 (Normalized, Figure 7), when the true class is 'helmet', it is correctly predicted 96% of the time, but 2% of 'helmet' instances are missed (predicted as background). When the true class is 'head', it is correctly predicted 95% of the time, with 3% missed. The 'person' class seems to be the most challenging, with only a small fraction of true 'person' instances being correctly identified, and many being missed or confused with the background.

The limited number of training epochs (7) is a significant factor. Deep learning models typically require more extensive training to converge and generalize effectively. The observed performance is likely an early-stage indication, and further training could lead to substantial improvements in recall for both models.

In terms of specific visual scenarios, the provided valida- tion batch images (Figures 17 through 22) show varying de- grees of success. For instance, in Figure 17 (YOLOv8), some occluded heads are missed in the prediction. In Figure 20 (YOLOv9), detections appear generally accurate for the pre- sented batch. A more detailed qualitative analysis across a wider range of challenging images (e.g., severe occlusion, poor lighting, very small helmets) would be necessary to draw firm conclusions about their robustness in specific adverse conditions.

## VI. CONCLUSION

This paper presented a comparative performance evaluation of YOLOv8 and YOLOv9 for safety helmet detection in construction environments, based on 7 epochs of training. YOLOv9 demonstrated a marginal but consistent improvement over YOLOv8 across key metrics, including mAP@0.5 (66.70% for YOLOv9 vs. 65.73% for YOLOv8) and overall F1-score (76.15% for YOLOv9 vs. 75.66% for YOLOv8).

The architectural innovations in YOLOv9, namely Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN), are likely responsible for its slightly superior learning capability ob- served even with limited training. Both models exhibited high precision, but their recall rates were found to be relatively low (around 0.62-0.63), indicating a significant number of missed helmet detections. This is a critical concern for safety applications and underscores the necessity for more extensive training to improve detection reliability.

Future work should prioritize training both models for a substantially larger number of epochs to allow for full convergence. A thorough error analysis is recommended to identify specific scenarios where detections are missed (e.g., occlusion, scale, lighting) to guide further model fine-tuning or dataset augmentation. Furthermore, evaluating

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, July 2025
Impact Factor: 7.67

inference speed (FPS) on hardware relevant to deployment scenarios is crucial for assessing real-time applicability. Ultimately, for practical deployment in construction safety, models must achieve higher recall rates

Future work should prioritize training both models for a substantially larger number of epochs to allow for full convergence. A thorough error analysis is recommended to identify specific scenarios where detections are missed (e.g., occlusion, scale, lighting) to guide further model fine-tuning or dataset augmentation. Furthermore, evaluating inference speed (FPS) on hardware relevant to deployment scenarios is crucial for assessing real-time applicability. Ultimately, for practical deployment in construction safety, models must achieve to ensure minimal missed violations

## REFERENCES

[1] M. Z. Shanti, C.-S. Cho, Y.-J. Byon, C. Y. Yeun, T.-Y. Kim, S.-K. Kim, and A. Altunaiji, "A Novel Implementation of an AI-Based Smart Construction Safety Inspection Protocol in the UAE," IEEE Access, vol. 9, pp. 166603- 166616, 2021.

[2] Y. Seth and M. Sivagami, "Enhanced YOLOv8 Object Detection Model for Construction Worker Safety Using Image Transformations," IEEE Access, vol. 13, pp. 10582-10594, 2025.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 779-788.

[4] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag- of-freebies sets new state-of-the-art for real-time object detectors," arXiv preprint arXiv:2207.02696, 2022.

[5] Ultralytics, "YOLOv8," GitHub. [Online]. Available: https://github.com/ ultralytics/ultralytics. (Accessed: Nov. 15, 2023).

[6] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," arXiv preprint arXiv:2402.13616, 2024.

[7] B. Lin, "Safety Helmet Detection Based on Improved YOLOv8," IEEE Access, vol. 12, pp. 28260-28272, 2024.

[8] R. Cheng, "A survey: Comparison between Convolutional Neural Network and YOLO in image identification," J. Phys.: Conf. Ser., vol. 1453, p. 012139, 2020.

[9] K. Shishodia, Manish, and V. Srivastava, "Performance Comparison of Yolo-V8 & Yolo-V9 on A Unified Traffic Sign Dataset," Grenze Interna- tional Journal of Engineering and Technology, vol. 10, no. 2, Jun. 2024.

[10] K. Han and X. Zeng, "Deep Learning-Based Workers Safety Helmet Wear- ing Detection on Construction Sites Using Multi-Scale Features," IEEE Access, vol. 10, pp. 718-729, 2022.

[11] J. Chen, J. Zhu, Z. Li, and X. Yang, "YOLOv7-WFD: A Novel Con- volutional Neural Network Model for Helmet Detection in High-Risk Workplaces," IEEE Access, vol. 11, pp. 113580-113592, 2023.

[12] H. Liang and S. Seo, "Automatic Detection of Construction Workers' Helmet Wear Based on Lightweight Deep Learning," Appl. Sci., vol. 12, no. 20, p. 10369, Oct. 2022.

[13] A. Haji et al., "Safety Helmet Detection Using YOLO V8," in 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN), Jun. 2023, pp. 22-26.