Deep Learning-Based Personal Protective Equipment Detection for Real-Time Healthcare Monitoring

1st Ludovica Beritelli
Dep. of Mathematics and Computer Science
University of Catania
Catania, Italy
ludovica.beritelli@phd.unict.it

2nd Stefano Antonio Amico
Department of Electrical, Electronic and Computer Engineering
University of Catania
Catania, Italy
stefano.amico@infoamico.com

3rd David Panebianco
Department of Electrical, Electronic and Computer Engineering
University of Catania
Catania, Italy
david.panebianco@phd.unict.it

4th Roberta Avanzato
Dep. of Electrical, Electronic and Computer Engineering
University of Catania
Catania, Italy
roberta.avanzato@unict.it

5th Francesco Beritelli
Dep. of Electrical, Electronic and Computer Engineering
University of Catania
Catania, Italy
francesco.beritelli@unict.it

***Abstract*—This work presents a deep learning-based framework for automatic Personal Protective Equipment (PPE) detection in healthcare environments. The proposed approach is based on a YOLO26n one-stage object detection architecture designed to achieve an effective trade-off between detection accuracy and real-time inference performance. A unified dataset was constructed by merging two publicly available healthcare-oriented datasets, resulting in 5,340 images and 13,308 annotated instances across four PPE categories: Coverall, Gloves, Goggles, and Mask. The model was trained using a transfer learning strategy and evaluated on an independent test set using standard COCO metrics along with Precision, Recall, and F1-score. Performance evaluation indicates that the proposed YOLO26n-based framework can accurately identify PPE items, yielding a mAP@0.5 of 0.944 together with Precision, Recall, and F1-score values of 0.912, 0.944, and 0.928, respectively. Additionally, the model achieves an average inference time of approximately 12.46 ms per image, demonstrating its suitability for real-time applications. Comparison with Faster R-CNN, YOLOv8n, and YOLO11n indicates that YOLO26n achieves the most favorable compromise between detection accuracy and inference efficiency. These results confirm the effectiveness of the proposed approach for real-time PPE monitoring in healthcare environments, where both accuracy and low-latency inference are essential requirements.**

***Index Terms*—Personal Protective Equipment (PPE), Object Detection, YOLO, Deep Learning, Healthcare Monitoring, Computer Vision.**

I. Introduction

The adoption of Personal Protective Equipment (PPE) is fundamental for reducing biological risks and preventing the spread of infectious diseases in healthcare environments. Devices such as face masks, gloves, gowns, and protective goggles constitute an essential barrier for both patients and medical personnel. However, ensuring their correct usage remains a challenging task. Manual monitoring procedures are inefficient, time-consuming, and difficult to scale in complex clinical settings characterized by the simultaneous presence of multiple operators.

To automate compliance monitoring, various approaches have been proposed, generally divided into sensor-based and vision-based systems. Approaches based on dedicated sensors, such as RFID systems or wearable devices, can operate reliably under different lighting conditions. However, they often require additional hardware installation, raise deployment costs, and provide limited information about whether PPE is worn correctly by healthcare operators [1]. Alternatively, early vision-based methods relied on handcrafted features (e.g., HOG, LBP) and classical classifiers [2]. Although computationally efficient, their performance degrades significantly in complex real-world scenarios involving occlusions, illumination changes, and viewpoint variations.

Consequently, the focus has shifted heavily toward Deep Learning. Deep learning methods based on convolutional neural networks have become the dominant solution for PPE recognition because they can learn task-specific visual features directly from data [3], [4]. Early deep learning pipelines relied on Regions of Interest (ROIs) classification, often employing lightweight architectures for edge deployment [5]. However, to overcome the need for preliminary segmentation, modern pipelines favor one-stage object detection architectures capable of simultaneous localization and classification. Models such as YOLO, SSD, and Faster R-CNN are widely used [6]–[8], demonstrating a strong trade-off between inference speed and detection accuracy in real-world scenarios for multi-class tasks.

Recent research also explores hybrid architectures combin-

ing object detection with human pose estimation to associate PPE with specific anatomical keypoints [1]. Although this improves robustness in crowded environments, achieving F1-scores up to 96% [5], it introduces additional computational overhead and remains sensitive to keypoint estimation accuracy under severe occlusions. Overall, finding the optimal balance between computational efficiency and detection robustness in dynamic healthcare settings remains an active research challenge.

To overcome the limitations highlighted in the previous discussion, this study investigates a real-time object detection solution for monitoring the use of Personal Protective Equipment (PPE) in healthcare settings. The proposed approach analyzes RGB images and identifies four key PPE categories, namely masks, gloves, goggles, and coveralls.

The main objectives and contributions of the research are summarized below:

- the development of an automated PPE detection framework capable of supporting real-time healthcare monitoring;
- the creation of a consolidated dataset obtained through the integration and harmonization of two publicly available PPE image collections;
- the validation of the proposed approach through extensive experimental analyses, including both quantitative performance assessment and qualitative evaluation in realistic scenarios.

The remainder of the paper is structured as follows. Section II outlines the adopted methodology, describing the dataset preparation process, the detection architecture, and the training configuration. Section III discusses the obtained results and provides a detailed performance evaluation. Finally, Section IV summarizes the main findings and presents potential directions for future research.

II. Proposed Method

A. Dataset Construction

To train and evaluate the proposed PPE detection system, two publicly available healthcare-oriented datasets were collected from the RoboFlow platform [9], [10] and merged into a unified dataset. A custom Python-based pipeline was developed to automatically integrate images and annotations while preserving only the object classes shared between the two datasets.

The selected datasets contain manually annotated RGB images acquired in realistic healthcare scenarios, including hospitals, clinical simulations, and medical laboratories. Consequently, the resulting dataset exhibits considerable variability in illumination conditions, camera viewpoints, image quality, background complexity, and number of subjects within each scene, improving the generalization capability of the trained model in real-world environments.

The dataset integration process involved class filtering, annotation remapping, and consistency verification. Initially, the common classes were identified through the corresponding `data.yaml` configuration files. Non-overlapping categories were discarded, while the remaining classes were reorganized according to a unified indexing strategy to ensure annotation consistency across all samples.

Subsequently, a dedicated preprocessing script automatically removed invalid annotations, updated class identifiers, and retained only images containing at least one valid PPE instance. To avoid filename conflicts, all samples were automatically renamed during the merging process. Finally, a new unified `data.yaml` file was generated, containing the updated class definitions and dataset structure.

The final dataset contains 5340 images and 13308 annotated object instances distributed across four PPE categories: *Coverall*, *Gloves*, *Goggles*, and *Mask*. Figure 1 shows representative examples extracted from the adopted datasets.

Fig. 1. Examples of images extracted from the selected PPE datasets.

Table I reports the distribution of annotated instances across training, validation, and test subsets. Although the dataset is slightly imbalanced, all categories contain a sufficient number of samples to support reliable training and evaluation procedures.

TABLE I
Distribution of annotated PPE instances across training, validation, and test sets.

Class	Training	Validation	Test	Total
Coverall	1966	367	278	2611
Gloves	2892	426	392	3710
Goggles	3048	515	454	4017
Mask	2238	413	319	2970
Total	10144	1721	1443	13308

The dataset was divided into three independent subsets: 4048 images for training (75.8%), 723 images for validation (13.5%), and 569 images for testing (10.7%). Prior to training, all images were resized to a fixed input resolution and normalized to improve training stability and ensure compatibility with the adopted Deep Learning architecture.

B. YOLO26-Based PPE Detection Framework

The proposed framework adopts the lightweight YOLO26n detector to perform PPE localization and classification in a single inference stage, enabling low-latency processing suitable for healthcare monitoring. Compared to traditional two-stage detectors, this design significantly reduces inference latency, making the architecture particularly suitable for real-time healthcare monitoring applications.

The YOLO26 architecture was adopted because it offers a practical balance between detection quality and inference speed, which is essential for real-time healthcare monitoring. The architecture incorporates multi-scale feature extraction mechanisms that improve the detection of PPE items appearing at different spatial resolutions, while enhanced feature aggregation strategies increase robustness under challenging conditions such as occlusions, cluttered backgrounds, and illumination variability.

Furthermore, the architecture improves discrimination capability between visually similar PPE categories, reducing false positives and increasing the reliability of the monitoring system in complex healthcare environments. The adopted framework is therefore particularly suitable for scenarios involving multiple healthcare operators simultaneously wearing different PPE items.

C. Experimental Setup

Training and inference procedures were implemented in Python 3.11.13 using the PyTorch framework (v2.10.0+cu128) together with the Ultralytics library (v8.4.33). Experimental activities were conducted on an NVIDIA A100-PCIE GPU equipped with 40 GB of VRAM.

The model was trained for 100 epochs using a batch size of 16 and an input image resolution of 640×640 pixels. Optimization was initialized with a learning rate of 0.01 and a warmup phase of 3 epochs to stabilize the early convergence process. The optimizer was automatically selected by the Ultralytics framework between SGD and AdamW according to the training configuration, while momentum and weight decay were set to 0.937 and 0.0005, respectively, to improve convergence stability and reduce overfitting.

The training objective combined three loss components: Box Loss for bounding box regression, Binary Cross-Entropy Classification Loss for category prediction, and Distribution Focal Loss (DFL) for localization refinement. The corresponding loss weights were set to 7.5, 0.5, and 1.5, respectively.

To accelerate convergence and improve generalization capability, a Transfer Learning strategy was adopted. Specifically, the network was initialized using pre-trained `yolo26n.pt` weights obtained from large-scale object detection datasets. Subsequently, fine-tuning was performed on the proposed PPE dataset, allowing the model to adapt its learned visual representations to the healthcare monitoring scenario.

III. Experimental Results

The proposed YOLO26-based PPE detection system was evaluated on the independent test set using the standard COCO evaluation protocol together with additional classification metrics, including Precision, Recall, and F1-score.

The model achieved a mAP@0.5 of 0.944 and a mAP@0.5:0.95 of 0.750, demonstrating strong localization and detection capabilities across all PPE categories. Table II reports the class-wise evaluation results.

TABLE II
Per-class Precision, Recall, and F1-score results on the test set.

Class	Precision	Recall	F1-score
Coverall	0.908	0.960	0.934
Gloves	0.882	0.931	0.906
Goggles	0.941	0.941	0.941
Mask	0.913	0.950	0.931

The obtained results indicate consistently high performance across all classes, with all F1-score values exceeding 0.90. In particular, the *Goggles* category achieved the best balance between Precision and Recall, obtaining the highest F1-score (0.941). The *Coverall* class achieved the highest Recall value (0.960), demonstrating strong capability in detecting most object instances belonging to this category.

The *Gloves* class exhibited slightly lower Precision compared to the other classes, suggesting the presence of a limited number of false positives. This behavior is likely related to the small size and visual similarity of gloves with surrounding scene elements under challenging illumination or occlusion conditions. Nevertheless, the class still achieved an F1-score above 0.90, confirming the robustness of the proposed approach.

A. Confusion Matrix Analysis

To further investigate the model behavior, the normalized confusion matrix shown in Figure 2 was analyzed.

The confusion matrix exhibits a strong concentration of values along the main diagonal, indicating high classification accuracy across all PPE categories. In particular, the model correctly identifies most *Coverall*, *Gloves*, *Goggles*, and *Mask* instances, with normalized true positive rates ranging from 0.93 to 0.96.

The analysis also reveals a limited number of false negatives associated with the *Background* class. These errors mainly occur in challenging scenarios involving small objects, partial occlusions, low illumination conditions, or cluttered scenes containing multiple subjects. Conversely, inter-class confusion remains limited, confirming the strong discriminative capability of the adopted architecture.

B. Inference and Qualitative Analysis

In addition to detection accuracy, the computational efficiency of the proposed model was evaluated through inference time analysis. Experiments conducted on the 569 test images revealed an average inference time of 12.46 ms per image,

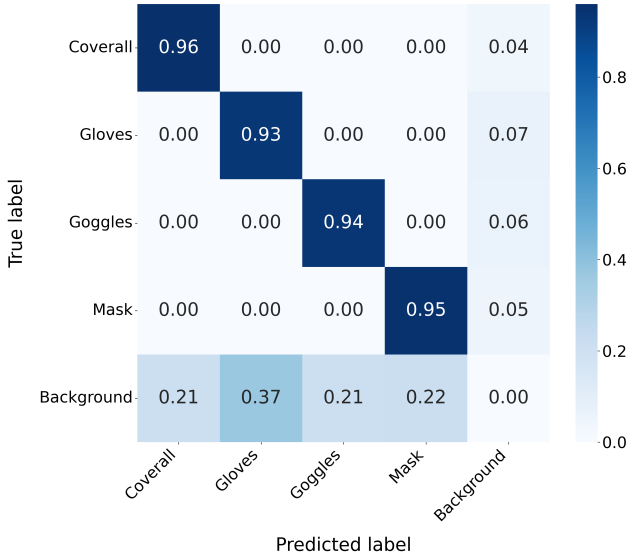

Fig. 2. Normalized confusion matrix.

with a standard deviation of 1.73 ms. Processing times ranged from a minimum of 10.63 ms to a maximum of 39.63 ms, demonstrating stable inference behavior.

These results confirm the suitability of the proposed system for real-time or near real-time healthcare monitoring applications, where low latency and reliable detection are both essential requirements.

A qualitative analysis was additionally performed on several test images. As shown in Figure 3, the model successfully detects multiple PPE items simultaneously, even in scenes characterized by multiple subjects, viewpoint variations, and complex backgrounds. Bounding boxes are generally well localized and consistent with the actual object positions.

Fig. 3. Examples of model inference on test images showing PPE detection through bounding boxes and predicted classes.

C. Comparison with Other Architectures

To further validate the proposed approach, YOLO26 was compared against other widely adopted object detection architectures, namely Faster R-CNN, YOLOv8n, and YOLO11n. Table III summarizes the obtained results.

TABLE III
COMPARISON BETWEEN DIFFERENT OBJECT DETECTION ARCHITECTURES ON THE PPE TEST SET.

Architecture	Precision	Recall	F1-score
Faster R-CNN	0.894	**0.959**	0.925
YOLOv8n	0.889	0.954	0.920
YOLO11n	0.888	0.947	0.917
YOLO26n	**0.912**	0.944	**0.928**

All evaluated architectures achieved strong performance on the PPE detection task. However, relevant differences emerge when jointly considering detection accuracy and computational efficiency.

Faster R-CNN achieved the highest Recall value, indicating strong capability in detecting most PPE instances. Nevertheless, its two-stage detection design introduces higher computational complexity and longer inference times, limiting its suitability for real-time healthcare monitoring applications.

Among the evaluated YOLO-based models, YOLO26n achieved the best overall balance, obtaining the highest Precision and F1-score while maintaining Recall values comparable to competing architectures. The improved Precision indicates a lower number of false positives, an important aspect in healthcare monitoring systems where incorrect detections may reduce system reliability.

Compared to Faster R-CNN, YOLOv8n and YOLO11n, YOLO26n shows competitive performance in terms of inference efficiency while providing improved detection quality. The results confirm that YOLOv8n is the fastest model in terms of average inference time, while YOLO26n maintains a very similar latency with only a marginal increase in computational cost. In contrast, Faster R-CNN exhibits significantly higher inference time and variability, further reinforcing its limitations for real-time deployment.

IV. CONCLUSION

This research focused on the design and performance assessment of an automated PPE detection system based on deep learning techniques for healthcare monitoring environments. A unified dataset was constructed by merging two publicly available datasets through an automated pipeline involving class filtering, annotation remapping, and consistency verification, ensuring a coherent and well-structured dataset for training and evaluation.

The proposed model, based on the YOLO26n architecture and implemented using the Ultralytics framework, was trained using a transfer learning strategy and optimized hyperparameters to improve convergence and generalization capability. The test-set evaluation confirmed the reliability of the detector, as every PPE class achieved Precision, Recall, and F1-score values above 90%, indicating accurate and balanced recognition performance. The analysis further confirms a robust classification capability, with limited inter-class confusion as evidenced by the confusion matrix.

Error analysis indicates that most failure cases are associated with false negatives, primarily caused by occlusions, small object scale, and challenging illumination conditions. Despite these limitations, the model exhibits stable behavior and strong robustness in realistic scenarios. In addition, the achieved inference time of approximately 12.5 ms per image confirms the suitability of the proposed approach for real-time PPE monitoring applications in healthcare environments.

Although the obtained results are highly satisfactory, several directions for future work can be identified. First, extending the dataset with additional samples and more diverse environmental conditions would further improve the model's generalization capability, particularly for challenging scenarios and under-represented cases. Second, the integration of advanced data augmentation techniques and optimization strategies could enhance robustness under adverse conditions such as severe occlusions and low visibility.

Future work will investigate lightweight transformer-based detectors and optimized edge deployment strategies to further reduce inference latency while preserving detection accuracy in hospital environments. Finally, a key research direction involves the deployment of the proposed model in real-world systems, such as edge devices or video surveillance platforms, enabling continuous and automated PPE compliance monitoring in healthcare facilities.

In conclusion, the experimental results indicate that the proposed detection pipeline can reliably recognize multiple PPE categories while maintaining inference times compatible with real-time deployment in healthcare environments.

REFERENCES

[1] Z. Pei, Q. Zhang, Y. Qi, Z. Wen, and Z. Zhang, "Identification of the normative use of medical protective equipment by fusion of object detection and keypoints detection," *Computer Methods and Programs in Biomedicine*, 2024.

[2] B. E. Mneymneh, M. Abbas, and H. Khoury, "Evaluation of computer vision techniques for automated hardhat detection in indoor construction safety applications," *Frontiers of Engineering Management*, vol. 5, no. 2, pp. 227–239, 2018.

[3] M. Ahmed, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Survey and performance analysis of deep learning based object detection in challenging environments," *Sensors*, vol. 21, no. 15, p. 5116, 2021.

[4] M. Trigka and E. Dritsas, "A comprehensive survey of machine learning techniques and models for object detection," *Sensors*, 2025.

[5] Y. Horesh, R. O. Rokach, Y. Kolben, and D. Nachman, "Real-time monitoring of personal protective equipment adherence using on-device artificial intelligence models," *Sensors*, vol. 25, no. 7, p. 2003, 2025.

[6] W. A. Shobaki and M. Milanova, "A comparative study of yolo, ssd, faster r-cnn, and more for optimized eye-gaze writing," *SCI*, vol. 7, no. 2, p. 47, 2025.

[7] A. K. et al., "Development and evolution of yolo: A survey," *Neurocomputing*, 2026.

[8] B. Çarklı Yavuz, "Scale-dependent performance analysis of yolo26 and yolov11 for ppe detection," *Electronics*, vol. 15, no. 6, p. 1146, 2026.

[9] CPPE, "Medical ppe dataset," https://universe.roboflow.com/cppe-nxmvw/medical-ppe-o6ot6 , may 2024, visited on 2026-05-29. [Online]. Available: https://universe.roboflow.com/cppe-nxmvw/medical-ppe-o6ot6

[10] O. Kawade, "Sppe_3 dataset," https://universe.roboflow.com/onkar-kawade/sppe_3 , jul 2024, visited on 2026-05-29. [Online]. Available: https://universe.roboflow.com/onkar-kawade/sppe_3