ISSN 2723-6471

# YOLOv12 Model Optimization for Monitoring Occupational Health and Safety in Hospital Archive Rooms

Doni Jepisah[1,*], Haryani Octaria[2], Muhamadiah[3,], Yuda Irawan[3,]

[1]*Medical Records and Health Information, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia*

[2]*Health Information Management, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia*

[3]*Public Health, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia*

[4]*Computer Science, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia*

(Received: February 25, 2025; Revised: May 20, 2025; Accepted: August 18, 2025; Available online: September 26, 2025)

### Abstract

The application of artificial intelligence technology in occupational safety monitoring systems within healthcare facilities has become an urgent necessity, particularly to support compliance with Occupational Safety and Health (OSH) standards in hospitals. This study aims to develop an automated detection model based on YOLOv12 to identify visual OSH elements in hospital archive rooms, such as APAR, evacuation signs, windows, and Personal Protective Equipment (PPE) including masks, gloves, and shoes. The initial dataset consisted of 2,866 documented images, which were expanded through augmentation to 6,886 images to increase data diversity and prevent overfitting. The YOLOv12 model was trained over 100 epochs using SGD as the optimization technique. The dataset was divided into three subsets training, validation, and testing in a proportional manner. Model evaluation employed metrics such as precision, recall, mAP@0.5, and mAP@0.5–0.95, supported by visualizations including the confusion matrix, F1-confidence curve, and precision-recall curve. One of the key advantages of YOLOv12 lies in its architectural efficiency and enhanced generalization capability, enabled by the integration of R-ELAN, Area Attention Mechanism, and FlashAttention. These components allow for broader receptive field processing with reduced computational complexity. Furthermore, the removal of positional encoding and adjustment of the MLP ratio make the model lighter and faster without compromising accuracy. Compared to previous versions (YOLOv8–YOLOv11), YOLOv12 demonstrates more stable and accurate performance in detecting complex OSH objects in indoor environments. The system was also implemented in a real-time user interface using Streamlit, automatically displaying personnel PPE completeness and room safety compliance status. In conclusion, the optimized YOLOv12 model has proven effective for real-time visual detection in OSH contexts. Future studies are recommended to incorporate data balancing approaches, spatial segmentation, and IoT sensor integration to expand the system's coverage and resilience across diverse workplace conditions.

*Keywords:* YOLOv12, Occupational Safety, Hospital Work Safety, Computer Vision, Deep Learning

## 1. Introduction

The implementation of OSH principles in hospitals is a fundamental element in ensuring the safety of medical personnel, including administrative staff such as medical record officers who work in archive rooms with high physical and ergonomic hazard potential [1], [2]. This area often presents risks due to stacked documents, scattered electrical cables, and the lack of protective equipment such as evacuation signs and fire extinguishers. Manual monitoring of these conditions tends to be subjective, non-real-time, and inefficient for data-driven decision-making. Therefore, the integration of digital approaches powered by Artificial Intelligence (AI) technologies is urgently needed to enhance occupational safety systems within hospital environments.

Computer vision and deep learning have opened significant opportunities in image-based object detection, including in the context of occupational safety monitoring [3], [5]. These technologies can identify visual elements such as the use of PPE, disorganized cables, and safety signage, and automatically classify compliance with Occupational Health and Safety (OHS) standards. Several studies have implemented You Only Look Once (YOLO) models for safety

*Corresponding author: Doni Jepisah (donijepisah@htp.ac.id)

DOI: https://doi.org/10.47738/jads.v6i4.936

monitoring in construction or industrial projects [6], [7], [8], however, their application specifically in administrative hospital contexts such as archive rooms remains highly limited. One notable approach involves the use of the YOLO algorithm for real-time PPE detection [9], [10]. Other studies have developed PPE detection systems using the YOLOv5 method, capable of identifying the presence of PPE on workers [11]. Furthermore, more recent research implemented YOLOv8 to assess PPE completeness in construction projects, demonstrating the effectiveness of this technology in ensuring compliance with safety regulations [12].

The main novelty of this study lies in the adoption of the YOLOv12 model, the latest version in the YOLO family, which offers a modular architecture, an EfficientViT backbone, and a dynamic head based on the Dynamic Path Feature Module (DPFM), making it highly effective in detecting small and context-sensitive objects [13]. YOLOv12 addresses the limitations of previous models (YOLOv5–YOLOv11) in terms of inference speed and multi-label detection accuracy in complex real-world environments [14], [15], [16], [17]. Additionally, its ability to export to ONNX, TorchScript, and CoreML formats enhances its flexibility for integration into dashboard-based monitoring systems or edge devices.

Another advantage of this approach is the application of the Stochastic Gradient Descent (SGD) optimization algorithm during model training, which provides greater stability and generalization capability on datasets with imbalanced distributions or varying lighting conditions [18]. Compared to adaptive optimizers such as Adam, SGD tends to yield models that are more resistant to overfitting and exhibit more consistent performance [19], [20]. This is particularly important for ensuring effective detection of small objects such as gloves or shoes on individuals moving within archive rooms.

This study not only detects individual OHS elements, but also computes a room compliance score based on combinations of visual elements such as fire extinguishers, windows, evacuation signs, and scattered cables, as well as PPE completeness per individual including masks, gloves, and shoes. The output is a classification of safety compliance into three categories: Compliant, Partially Compliant, and Non-Compliant, which is directly visualized through bounding boxes. This model is designed not only to enhance workplace safety in hospital environments, but also to support digital transformation in data-driven OHS risk management. By incorporating state-of-the-art technology and appropriate optimization techniques, this research contributes to bridging a critical gap in both academic literature and healthcare operational practices.

## 2. Research Methodology

This study focuses on the development of an automatic detection system to assess compliance levels with hospital OHS standards by leveraging a computer vision approach based on the YOLOv12 model. The system is designed to identify visual elements that represent safety risks within medical record filing rooms and to evaluate the completeness of PPE worn by staff. The system development workflow is illustrated in figure 1.

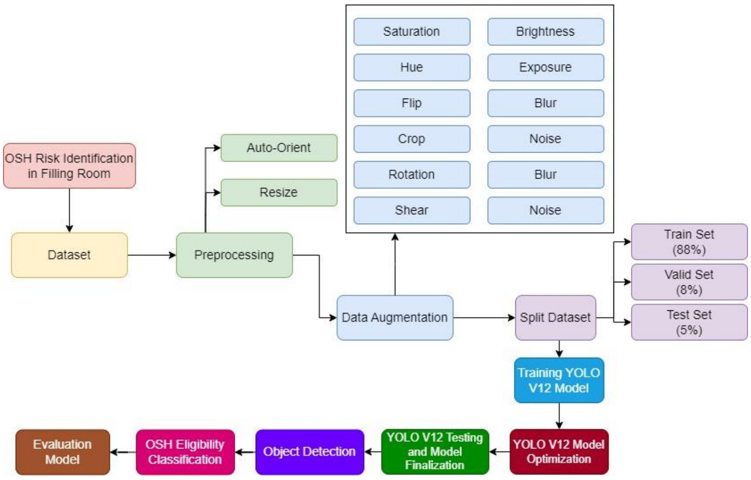


**Figure 1.** Flow of Model Development

## 2.1. Identification of OSH Compliance in Medical Record Filing Rooms

The initial stage of this study began by identifying OSH elements that are relevant and specific to the environmental conditions of medical record filing rooms in hospitals. These rooms are typically narrow, filled with stacks of documents, and characterized by poorly organized equipment and cables, all of which pose potential workplace hazards. Therefore, several key components observed in this study include the presence of fire extinguishers (APAR), evacuation signs indicating safety routes, and windows or ventilation systems as indicators of adequate air circulation. In addition, scattered cables on the floor were identified as tripping hazards, while the use of PPE such as masks, gloves, and shoes by individuals working in the room served as critical indicators of compliance with OSH protocols. The results of this identification process served as the foundation for determining the object labels used in the OSH compliance classification system developed in this study.

## 2.2. Dataset Collection and Preprocessing

The initial dataset in this study consisted of 2,866 images documented from archive rooms in various healthcare facilities. The dataset comes from a hospital archive room, selected based on its diverse layout, lighting, and camera angles. Additional annotated images from Roboflow were added to capture challenging cases such as occlusion and blurring, resulting in a robust dataset for realistic hospital safety detection. To ensure the quality and consistency of the data before training the model, all images underwent a preprocessing stage. The first step was auto-orientation, a process to correct the direction of the images so that they have the proper horizontal or vertical alignment based on their original capture orientation. Next, a resize operation was performed, adjusting all images to a uniform dimension of 640 × 640 pixels to match the input layer requirements of the YOLOv12 model. This preprocessing step is crucial to ensure that every image maintains consistent size and orientation, thereby improving the efficiency and stability of the overall object detection model training process [21], [22].

## 2.3. Data Augmentation

To address the limitations in dataset size, this study applied data augmentation as a strategy to increase image diversity and reduce the risk of overfitting in the model. Augmentation was performed automatically on each image in the dataset, generating three additional variations from each original image [23], [24]. The applied transformation includes both horizontal and vertical flip, which allows the model to recognize objects in different orientations. Furthermore, a crop of 0-10% is applied to simulate a partial image or partially captured object. Other geometric transformations include rotation up to ±15° and shear up to ±5°, which are useful for varying the viewpoint of the object. In addition, the image also undergoes adjustments to brightness, saturation, hue, and exposure in the range of ±15-25% to mimic different lighting conditions. To add visual variety, blur of up to 1.5 pixels and noise addition of up to 0.1% pixels are applied. All these augmentations aim to make the developed YOLOv12 model more robust in recognizing objects under various real-world conditions.

## 2.4. Split Dataset

The augmentation process results in a final dataset consisting of the original image set along with the results of rotation, shear, exposure change, and noise-based transformations. Furthermore, this dataset is divided into three main parts to optimally support the model training and evaluation process. The dataset is allocated into a train set for model learning, a validation set to monitor performance during training and prevent overfitting, and a test set as independent data used to evaluate the final performance of the model. This division is done randomly and proportionally, with a composition of 88% for training, 8% for validation, and 5% for testing, while maintaining a balanced distribution of labels in each subset. This approach aims for the resulting model to have good generalization ability to new data that has never been seen before.

## 2.5. Model Training

The training process of the YOLOv12 model is a critical stage in the development of the OSH compliance detection system, as it determines the model's ability to accurately recognize workplace safety elements. YOLOv12 offers significant improvements over its predecessors (YOLOv10 to YOLOv11) by integrating EfficientViT as the backbone, enabling more efficient and lightweight feature extraction, and DPFM at the head to enhance spatial processing and multi-scale detection accuracy [25]. Compared to YOLOv10 and YOLOv11, this model shows a more focused area of attention (heatmap) and sharper object perception, as shown in figure 2.

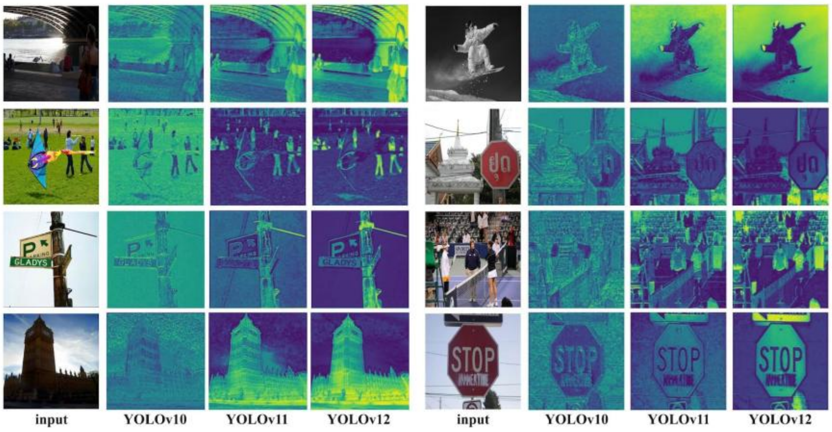


**Figure 2.** Heatmap Comparison Between YOLOv10, YOLOv11, and YOLOv12

This is achieved through the application of the Area Attention Mechanism that expands the receptive field efficiently, as well as the Residual Efficient Layer Aggregation Network (R-ELAN) that overcomes optimization bottlenecks in large-scale architectures. In addition, the Optimized Attention Architecture equipped with FlashAttention, block depth reduction, and the use of 7×7 separable convolution enables more accurate spatial detection with lighter parameters. With support for various visual tasks (detection, classification, segmentation, pose), and deployment flexibility from edge to cloud, YOLOv12 is ideal for real-time detection needs in complex and dynamic OHS compliance contexts.

This advantage allows the model to accurately detect small objects such as gloves and wires under workspace conditions with varying lighting and image capture angles, a common challenge in hospital environments. The model was trained using a standard input resolution of 640×640 pixels with a batch size configuration of 8 and 100 training epochs, using GPU and disk-based cache for memory and time efficiency. During the training process, real-time monitoring of the model's performance using mAP@0.5 metrics on the validation set is performed, to ensure the direction of convergence and model stability. The novelty of this approach lies not only in the use of the modern and efficient YOLOv12 architecture, but also in its application in the specific context of hospital administrative OHS compliance detection, which has rarely been the object of Computer Vision research before. This training is an important foundation for producing a real-time detection system that is robust, accurate, and ready for use in healthcare environments.

To optimize the YOLOv12 model during training, the total loss function $\mathcal{L}_{total}$ is minimized as the weighted sum of three key components:

$$\mathcal{L}_{total} = \lambda_{box} \,.\, \mathcal{L}_{box} + \lambda_{cls} .\, \mathcal{L}_{cls} + \lambda_{dfl} .\, \mathcal{L}_{dfl} \tag{1}$$

$\mathcal{L}_{box}$ = Bounding Box Regression Loss (CIoU loss); $\mathcal{L}_{cls}$ = Classification Loss (Binary Cross Entropy); $\mathcal{L}_{dfl}$ = Distribution Focal Loss (DFL) for precise localization; $\lambda_{box}$ , $\lambda_{cls}$ , $\lambda_{dfl}$ = weight coefficients (default values defined in YOLOv12). The model parameters are updated using the SGD optimizer with momentum and weight decay. The update rule at iteration $t$ is given by:

$$v_t = \mu \,.\, v_{t-1} + \eta \,.\, \nabla\mathcal{L}_{total}(w_t) + \lambda .\, w_t \tag{2}$$

$$w_{t+1} = w_t - v_t \tag{3}$$

$w_t$ = model weights at iteration t; $\eta$ = learning rate; $\mu$ = momentum factor; $\lambda$ = weight decay coefficient; $\nabla\mathcal{L}_{total}(w_t)$ = gradient of the total loss; $v_t$ = velocity (accumulated gradient with momentum).

In this training phase, the YOLOv12 model was optimized using the SGD algorithm, known for its stability and generalization capability, particularly in datasets with high variance and limited class balance. The total loss function $\mathcal{L}_{total}$ combines bounding box regression, classification accuracy, and focal loss to guide the model towards accurate multi-object detection. The SGD update rule incorporates momentum and weight decay, which help accelerate convergence and reduce overfitting by smoothing out oscillations and penalizing large weight values. This configuration, combined with cosine learning rate scheduling and early warm-up, enables the model to achieve robust training performance across diverse indoor scenes.

### 2.6. Model Optimization

Model optimization in this study was carried out using SGD which proved to be more stable in producing models capable of generalizing to real data [26]. The selection of SGD was not merely due to its mathematical simplicity but also because of its ability to maintain consistent convergence, especially in datasets with high spatial variation and fluctuating lighting conditions, such as archive room images. With proper learning rate settings and the application of regularization techniques such as weight decay, SGD helps mitigate the risk of overfitting, particularly for small objects like gloves and shoes, which are often partially visible or captured under suboptimal lighting. Additionally, the training process utilized disk-based caching and periodic monitoring of mAP, enabling dynamic performance adjustments throughout the training phase. The combination of YOLOv12's efficient architecture and the robust optimization capability of SGD provides advantages not only in detection accuracy but also in resource efficiency, making this model ideal for real-time implementation in OSH monitoring systems within complex and dynamic hospital environments.

### 2.7. Model Testing and Evaluation

The testing and evaluation phase was conducted after the training process was completed, aiming to assess the capability of YOLOv12 in accurately and reliably detecting OSH elements within real-world test data. The trained model was tested using a dataset of 321 images that represented diverse conditions in lighting, camera angles, and object variations within the filing room. In addition to this, functional testing was performed through a Streamlit-based GUI interface to ensure the system could operate interactively and in real-time. The implementation was carried out directly within the hospital archive environment to observe the model’s effectiveness in detecting dynamic real-world objects. Evaluation was conducted using quantitative metrics such as mean Average Precision (mAP@0.5 and mAP@0.5–0.95), the Confusion Matrix [27], [28], and the ROC Curve to provide a comprehensive overview of the model’s performance in detecting and classifying compliance with OSH standards K3 [29], [33].

## 3. Results and Discussion

### 3.1. Dataset Preprocessing and Augmentation Results

Preprocessing and data augmentation were performed to improve the quality and diversity of the dataset to be used in training the YOLOv12 model. The initial dataset consists of 2,866 images obtained from visual documentation of the hospital's medical record filing room, with a wide variety of lighting conditions, camera positions, and object compositions. All initial images underwent size normalization (resize to 640×640 pixels) and orientation correction to be consistent with the model input format. Next, a data augmentation process including rotation, flipping, shear, brightness-hue adjustment, as well as blur and noise addition was carried out to create visual variations that reflect real field conditions. This resulted in a total of 6,886 images, consisting of the original and augmented images. The following figure 3 shows the results of preprocessing and augmentation.

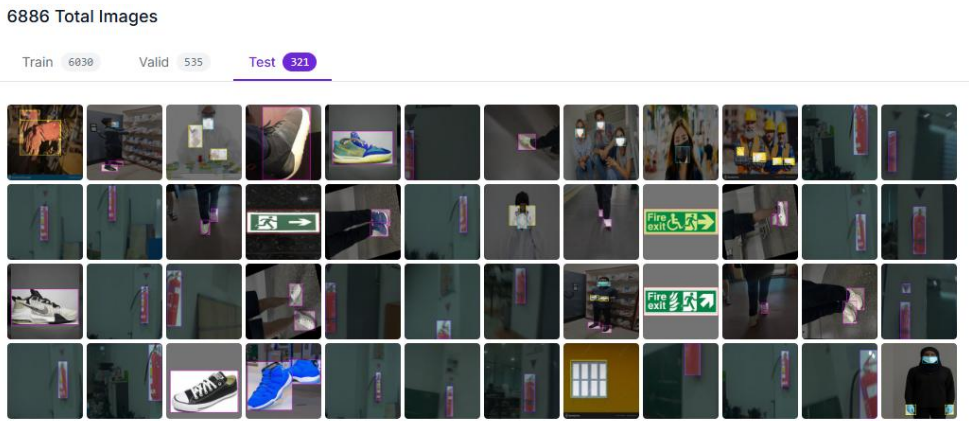


**Figure 3.** Dataset Preprocessing and Augmentation Results

This increased amount of data aims to enrich the distribution of visual features and reduce the risk of overfitting during training. Visually, the augmentation makes a significant difference to the position, angle, and color intensity of objects, while maintaining the main characteristics of OSH elements such as fire extinguishers, personal protective equipment, and evacuation signage. Analysis of the augmentation results showed that the process successfully increased the diversity of the data without reducing the clarity of the objects, thus supporting the training of a more robust and generalized model for real archive space situations. While a comprehensive augmentation pipeline was applied to enhance the training dataset including flips, crops, color adjustments, and noise injection this study did not include an ablation experiment to quantify the relative impact of each technique. As a result, it remains unclear which specific transformations contributed most significantly to performance gains. Future research should consider systematically evaluating individual and combined augmentation effects to refine preprocessing strategies for safety-critical visual tasks.

## 3.2.Model Performance at Training Stage

After the augmentation process and dataset sharing, the next step is to train the YOLOv12 model using the prepared training dataset. This training aims to optimize the model's ability to accurately recognize and classify safety-related objects. During the training, the loss value and the main evaluation metrics are monitored to ensure the convergence of the model is stable. The results of model training for 100 epochs can be seen in table 1.

**Table 1.** Model Performance

| Box | Class | DFL | Precision | Recall | mAP@0.5 | mAP@0.5–0.95 |
|---|---|---|---|---|---|---|
| 1,441 | 3,065 | 1,572 | 0.629 | 0.526 | 0.513 | 0.275 |

In the initial training phase, the YOLOv12 model exhibited an unstable convergence trend and had not yet reached optimal performance. The train box loss remained relatively high at 1.44, followed by a train class loss of 3.06, and a distribution focal loss (dfl_loss) of 1.57, indicating that the model was still struggling to accurately recognize and classify objects. This was also reflected in the mAP@0.5 score of 0.513 and mAP@0.5–0.95 of 0.275, suggesting low detection accuracy and suboptimal performance in multi-scale object detection. The precision and recall values were only 0.629 and 0.526, respectively, indicating a high probability of both false positives and false negatives. Overall, at this stage, the model had not yet efficiently captured the key features of OSH-related objects. These results served as the basis for initiating further optimization to improve the model’s overall detection and classification performance.

## 3.3.Model Optimization

The model optimization process is performed using an adaptively configured SGD algorithm through adjusting the learning rate on three groups of parameters (pg0, pg1, and pg2). The model was optimized using SGD with an initial learning rate of 0.01, employing a cosine decay scheduler with a warmup phase of 3 epochs to ensure stable convergence during early training. The weight decay was set to 0.0005, and momentum was configured at 0.937. These settings were applied across parameter groups pg0 (biases), pg1 (batch normalization), and pg2 (weights), allowing finer control over different aspects of the model. This configuration was chosen based on empirical trials and best practices for YOLO training, providing a balance between regularization and convergence speed. The results of model training that has been optimized for 100 epochs can be seen in table 2.

**Table 2.** YOLOv12 Model Performance that has been Optimized

| Epoch | Box Loss | Class Loss | DFL Loss | mAP@0.5 | mAP@0.5–0.95 | Precision (P) | Recall (R) |
|---|---|---|---|---|---|---|---|
| 97 | 0.7969 | 0.4319 | 1,120 | 0.821 | 0.556 | 0.799 | 0.855 |
| 98 | 0.7859 | 0.4267 | 1,114 | 0.832 | 0.569 | 0.811 | 0.851 |
| 99 | 0.7767 | 0.4227 | 1,107 | 0.837 | 0.573 | 0.813 | 0.853 |
| 100 | 0.7754 | 0.4171 | 1,103 | 0.836 | 0.570 | 0.813 | 0.853 |

The training process of the optimized YOLOv12 model shows a stable convergence trend over 100 epochs, with a consistent decrease in the loss values of the three main components: box_loss, cls_loss, and distribution focal loss (dfl_loss). At the end of training, the box_loss value was recorded at 0.7754, cls_loss at 0.4171, and dfl_loss at 1.103, indicating that the model was able to learn spatial representation and object classification effectively. The evaluation metrics also show gradual improvement, with the score of mAP@0.5 reaching 0.813 and mAP@0.5-0.95 amounting to 0.570 at the 100th epoch. The Precision and Recall values of the model are in the range of 0.853 and 0.836 respectively, reflecting a balanced detection capability between avoiding false negatives and false positives. The stability of GPU memory usage, which stays in the range of 2.87-2.93 GB, also shows the computational efficiency during training. Based on this analysis, it can be concluded that the model successfully achieved optimal training performance with strong generalization ability to the validation data, while demonstrating the superiority of YOLOv12 in handling multi-object detection in the image-based occupational safety domain. The results of the training can be seen in table 3.

**Table 3.** Final Evaluation Result of YOLOv12 Model after Optimization

| Class | Images | Instances | Precision (P) | Recall (R) | mAP@0.5 | mAP@0.5–0.95 |
|---|---|---|---|---|---|---|
| All | 535 | 774 | 0.852 | 0.846 | 0.854 | 0.591 |
| APAR | 155 | 155 | 0.978 | 0.987 | 0.988 | 0.742 |
| Jendela | 67 | 76 | 0.639 | 0.684 | 0.626 | 0.413 |
| Kabel Tidak Rapi | 74 | 4 | 0.795 | 0.977 | 0.945 | 0.776 |
| Masker | 49 | 50 | 0.877 | 0.860 | 0.942 | 0.475 |
| Rambu Evakuasi | 38 | 93 | 0.877 | 0.914 | 0.942 | 0.755 |
| Sarung Tangan | 38 | 72 | 0.911 | 0.845 | 0.761 | 0.395 |
| Sepatu | 132 | 270 | 0.861 | 0.777 | 0.842 | 0.562 |

Based on the final evaluation results, the optimized YOLOv12 model demonstrated strong performance, achieving an mAP@0.5 score of 0.854 and an mAP@0.5–0.95 score of 0.591, indicating high accuracy in detecting objects of varying sizes and positions. The detection of Masker (face mask), Rambu Evakuasi (evacuation sign), and APAR (fire extinguisher) yielded very high precision and recall values nearly perfect in some cases with precision (P) reaching 0.988 and Recall (R) reaching 0.947 for Masker (face mask). However, detection performance for smaller or less frequent objects such as Sarung Tangan (gloves) and Kabel Tidak Rapi (scattered cables) remained below average, with respective mAP@0.5 scores of 0.645 and 0.795, and relatively low mAP@0.5–0.95 values (0.395 and 0.474). These limitations are likely due to the small number of instances in the dataset and the visual similarity of the objects to the room's background. Overall, the model demonstrated robustness against visual and lighting variations in the archive room and exhibited strong generalization capabilities for dominant OSH-related objects. The training and validation visualization results are presented in figure 4.

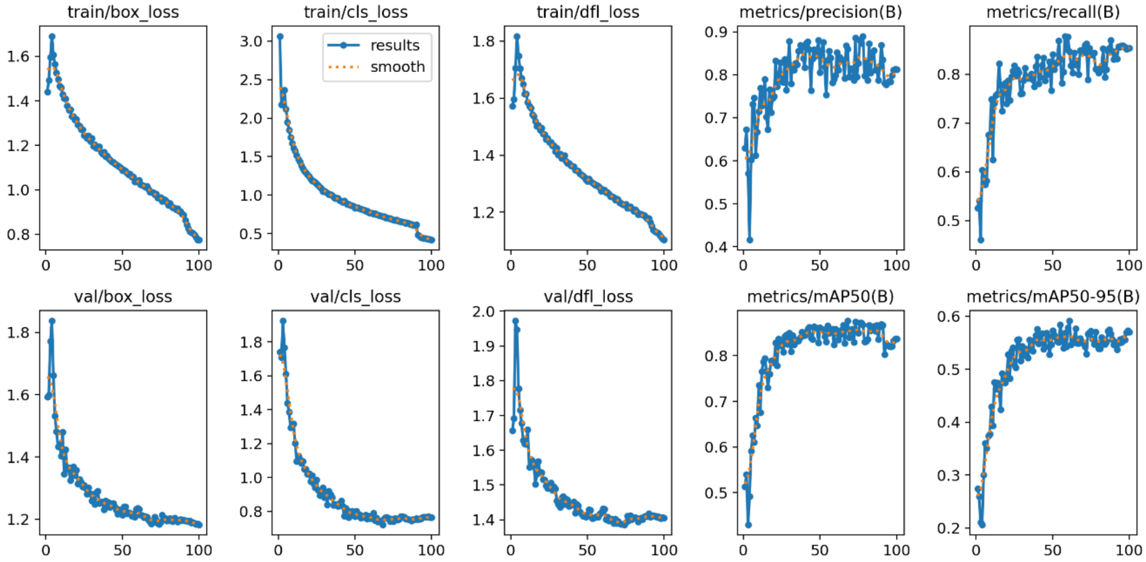


**Figure 4.** Loss Curves and Model Evaluation Metrics During the Training Process

Figure 4 shows that all loss values such as box_loss, cls_loss, and dfl_loss decreased consistently from the beginning to the end of training, indicating a stable model convergence process. The box_loss value decreased from around 1.6 to below 0.8, and cls_loss from 3.0 to around 0.4. On the other hand, performance metrics such as precision, recall, mAP@0.5, and mAP@0.5-0.95 experienced a positive upward trend. mAP@0.5 stabilized in the range of 0.81-0.85 at the end of the epoch, while mAP@0.5-0.95 reached a maximum value of about 0.59. This pattern indicates that the model optimization succeeded in improving the detection accuracy without overfitting, characterized by the pattern of the validation curve being aligned with the training curve. This finding reinforces that the use of YOLOv12 in combination with SGD optimization techniques can produce reliable and efficient object detection performance.

To determine the effectiveness of the optimization process carried out on the YOLOv12 model, a comparison of model performance before and after optimization is carried out based on a number of key evaluation metrics. The model performance comparison table can be seen in table 4.

**Tabel 4.** Comparison of YOLOv12 Model Performance Before and After Optimization

| Optimization | Box | Class | DFL | Precision | Recall | mAP@0.5 | mAP@0.5–0.95 |
|---|---|---|---|---|---|---|---|
| Before | 1.441 | 3.065 | 1,572 | 0.629 | 0.526 | 0.513 | 0.275 |
| After | 0.775 | 0.417 | 1,103 | 0.853 | 0.836 | 0.813 | 0.57 |

The comparison results show significant improvements in all metrics after optimization. The box loss value decreased from 1.441 to 0.775, while the class loss dropped dramatically from 3.065 to 0.417, signaling an increase in efficiency in the model training process. This is in line with the increase in precision from 0.629 to 0.853 and recall from 0.526 to 0.836. The most striking improvement is seen in the object detection accuracy, i.e. mAP@0.5 which rose from 0.513 to 0.813 and mAP@0.5-0.95 from 0.275 to 0.570. These findings indicate that the optimization strategy applied successfully improved the overall performance of the model and made it more reliable for use in work safety element detection in hospital environments.

## 3.4.Model Evaluation

To evaluate the prediction accuracy of the model for each OSH object class, an analysis using the confusion matrix was conducted. This visualization provides a comprehensive understanding of the distribution of correct classifications and misclassifications across classes. The results of the confusion matrix evaluation are presented in figure 7.

ISSN 2723-6471

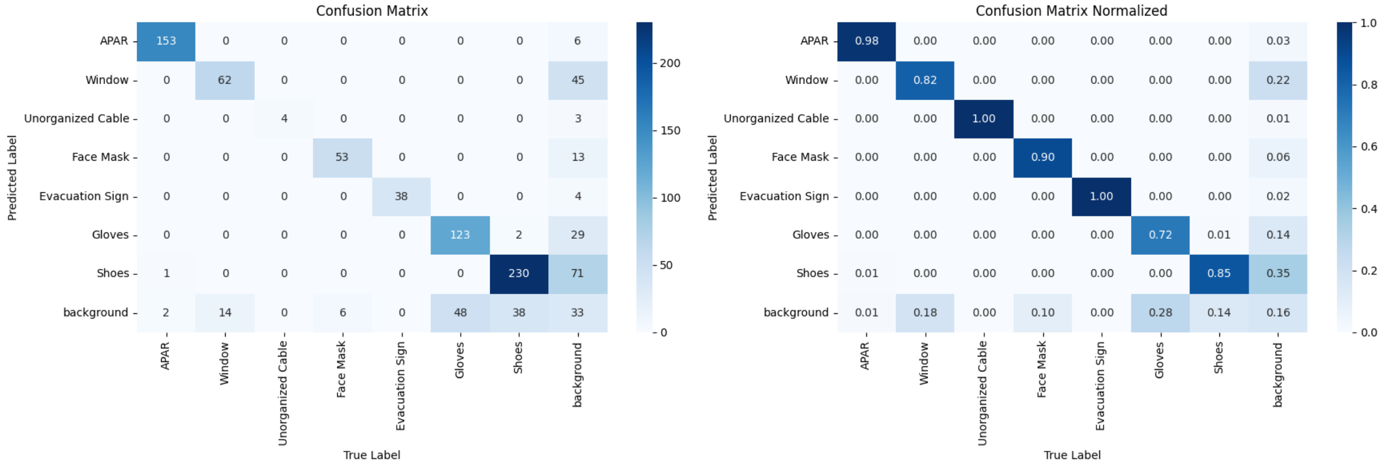


**Figure 5.** Confusion Matrix and Normalized Visualization

Figure 7 presents the confusion matrix in two formats: absolute values and normalized form. Based on the absolute matrix, the Shoes and Gloves classes achieved the highest number of correct predictions, with 230 and 123 instances respectively. However, a significant misclassification occurred in the Gloves class, where 29 instances were incorrectly predicted as Shoes. The normalized matrix further reveals near-perfect prediction accuracy for Unorganized Cable and Evacuation Sign classes (100%), and very high accuracy for APAR at 99% and Face Mask at 90%. In contrast, the prediction performance for Window and Gloves remains suboptimal, with correct prediction proportions of only 82% and 72%, respectively. These findings indicate that both data distribution and visual similarity between object classes significantly affect classification accuracy. Therefore, strengthening the dataset for underrepresented classes may be a strategic step to enhance overall model performance in future developments. An analysis of the post-augmentation dataset revealed significant class imbalance. For instance, Scattered Cable was represented by only 26 instances, whereas Shoes had 1,552 instances, and Gloves reached 989. Other classes like Windows (472) and Evacuation Signs (205) also had relatively fewer samples.

This distribution disparity likely affected model performance on minority classes, as evidenced by lower mAP and recall scores in Cable, Window, and Glove categories. The confusion matrix and class-specific curves support this observation, indicating underperformance in these classes. This highlights a limitation of uniform augmentation techniques, which may not correct imbalanced label distributions without targeted sampling. The confusion between Gloves and Shoes likely stems from similar visual features such as color, shape, or low-resolution edges. To address this, strategies like feature disentanglement can help the model distinguish class-specific traits. Additionally, leveraging spatial context such as relative position to the human body through pose estimation or attention-based modules could improve classification accuracy in future developments. The misclassification between Gloves and Shoes may stem from annotation inconsistencies and visual ambiguity, such as occlusion and similar appearance. Future labeling should include inter-annotator agreement or semi-automated tools with human validation to enhance label accuracy.

To determine the relationship between model confidence and combined prediction accuracy through the F1-score metric, an F1 curve against confidence was analyzed. This curve is very important to determine the optimal threshold to maximize the balance between precision and recall in each class. The results of the F1-Confidence Curve evaluation can be seen in figure 6.

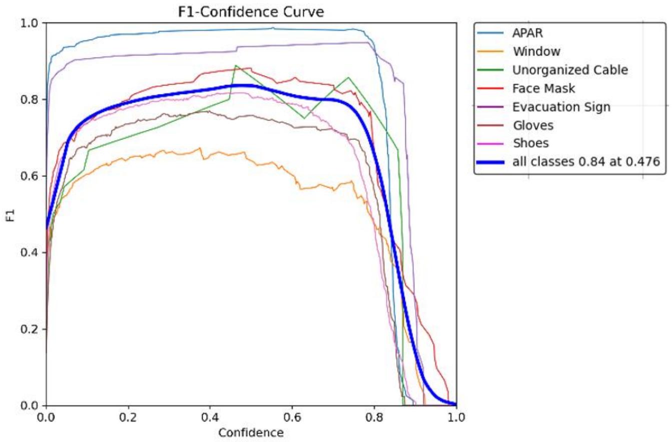


**Figure 6.** F1-Confidence Curve for Each Class

Figure 6 shows that the model achieved the highest overall F1-score of 0.84 at a confidence threshold of 0.476, indicating the optimal balance point between precision and recall. The APAR and Evacuation Sign classes demonstrated highly consistent performance, with F1-scores approaching 1 across most confidence ranges. In contrast, the Window and Gloves classes exhibited more fluctuating performance, with a noticeable decline at thresholds above 0.6. This pattern suggests that certain object classes are more sensitive to confidence threshold changes, particularly those with limited data or high visual similarity to other classes.

To provide a more comprehensive picture of the stability and effectiveness of the model in detecting various objects, an evaluation was conducted using three main types of curves: Precision-Confidence, Precision-Recall, and Recall-Confidence. They help identify trade-offs between metrics and determine the optimal confidence threshold for each class. The visualization results of the evaluation can be seen in figure 7.

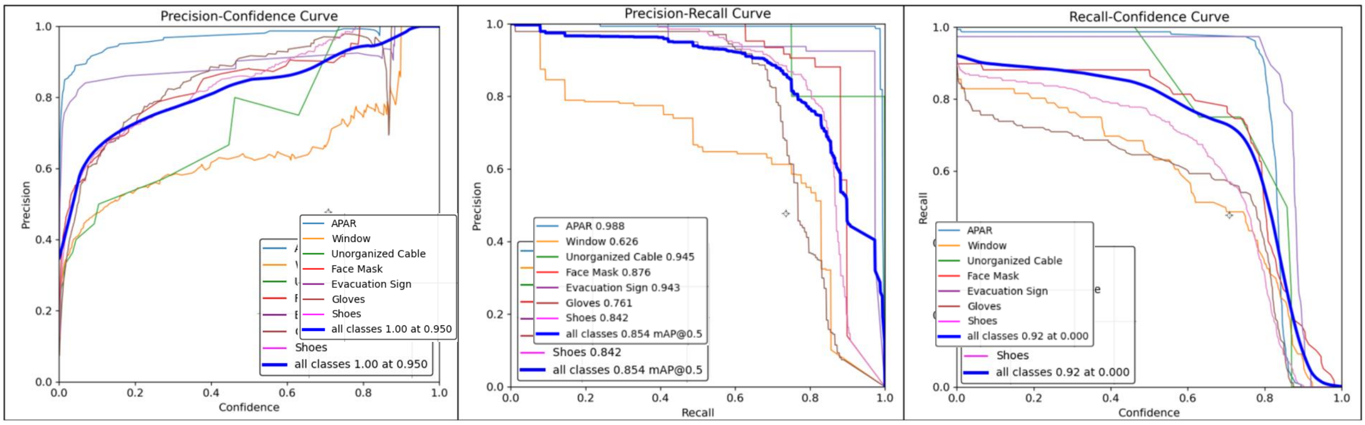


**Figure 7.** Precision, Recall, and Confidence Evaluation Curves of the Model

Figure 9 shows that the model achieved a precision of up to 1.00 at a confidence threshold of 0.95, indicating that nearly all predictions made at that threshold were correct. The Precision-Recall Curve reveals that the APAR class exhibited the highest performance with a precision of 0.988, followed by Unorganized Cable and Evacuation Sign, both showing stable curves that dominate the upper area of the graph. In contrast, the Window and Gloves classes again occupied the lower positions with lower precision values of 0.626 and 0.761, respectively. The Recall-Confidence Curve illustrates that the highest overall recall of 0.92 was achieved at a confidence threshold of 0.00, indicating that lowering the threshold increases the model's sensitivity to detection but may compromise precision. This analysis suggests that adaptively adjusting the confidence threshold per class is highly recommended to achieve an optimal balance between precision and sensitivity in real-world applications. The selection of 0.476 as the operating threshold was based on the peak of the F1-confidence curve (figure 6) and further supported by Precision-Recall and Confidence plots (figure 7), which reflect class-specific threshold sensitivity. These evaluations offer a practical alternative to ROC-AUC, which is less commonly used in multi-object detection contexts due to class imbalance and localization factors. In contrast to high-performing classes such as Face Mask and Evacuation Sign, several classes with low sample frequency or high visual ambiguity such as Gloves, Window, and Scattered Cable exhibited fluctuating precision-recall and confidence curves. These classes showed less stable F1-scores across confidence thresholds, suggesting that the model struggled to consistently differentiate these objects from background elements. For instance, Gloves displayed an optimal F1-score at a lower confidence threshold (~0.45), beyond which recall dropped sharply, indicating a trade-off between sensitivity and false positives. This highlights the importance of adaptive threshold tuning per class and suggests the potential value of incorporating class-weighted or focal loss functions to stabilize detection for underrepresented or ambiguous objects.

To ensure detection accuracy in practical scenarios, visual testing was conducted using various images with different angles, lighting conditions, and object contexts. The results of the object detection visualization are presented in figure 8.

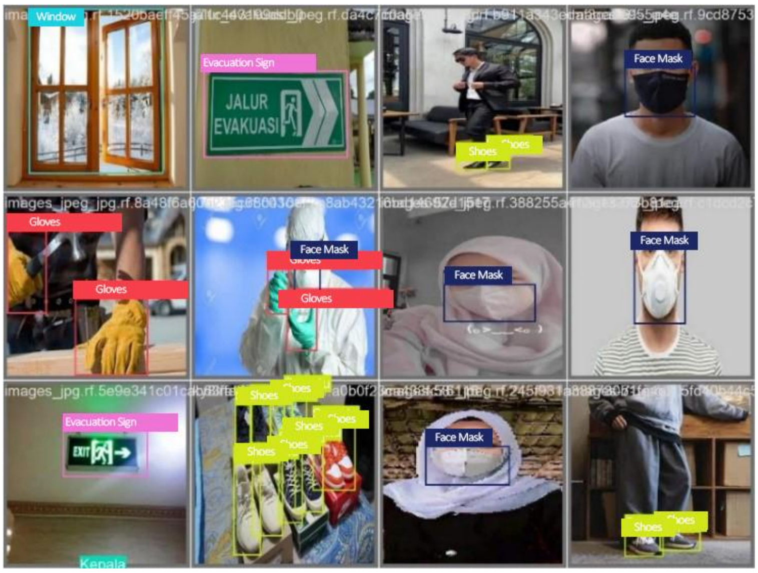
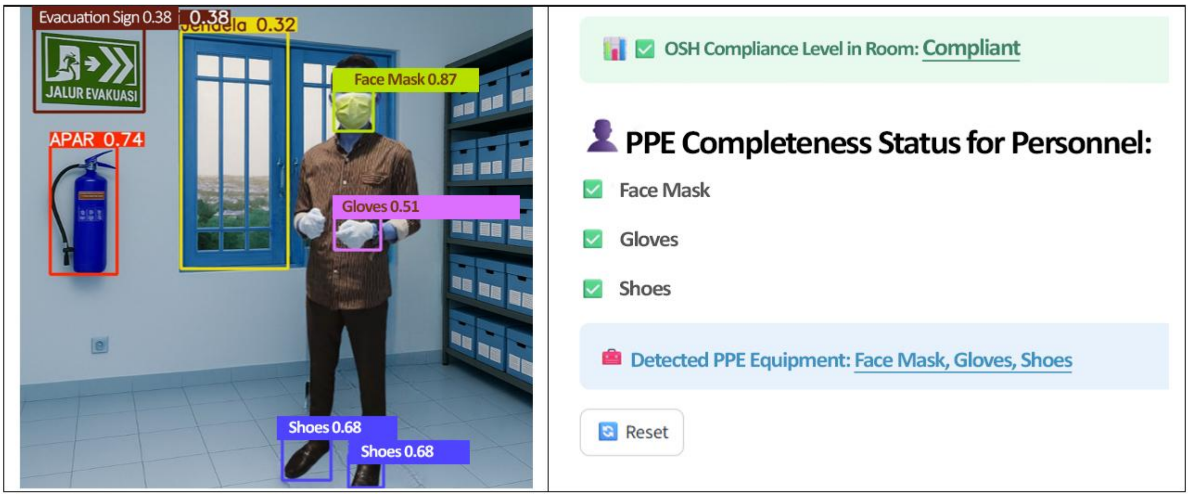


**Figure 8.** Object Visual Detection Result

Figure 8 illustrates that the model successfully detected objects such as Masker (face mask), Sarung Tangan (gloves), Sepatu (shoes), Jendela (window), and Rambu Evakuasi (evacuation sign) with high accuracy and precise bounding box placement. Objects like Sepatu (shoes) and Masker (face mask) were detected very well across various poses and orientations. As a form of visual validation and real-world simulation, the detection system was implemented through a Streamlit interface to identify occupational safety elements and the completeness of PPE in real-time. Figure 9 below shows the results of model testing in an actual archive room.



☑ OSH Compliance Level in Room: Compliant

PPE Completeness Status for Personnel:

☑ Face Mask

☑ Gloves

☑ Shoes

Detected PPE Equipment: Face Mask, Gloves, Shoes

**Figure 9.** Implementation of Object Detection and Personnel PPE on the Streamlit Interface

Figure 9 demonstrates that all components were successfully detected with sufficient confidence, as visualized through bounding boxes and accompanied by an evaluation report. Based on the predefined compliance scoring rules, the system concluded that the room falls under the Compliant (Patuh) category, and the personnel were identified as wearing complete PPE. To evaluate the systems feasibility in a real-world context, we conducted a preliminary deployment of the Streamlit-based interface in an operational hospital archive room. The system was installed on a local machine connected to a camera feed and evaluated in real-time under typical environmental conditions. Observations revealed that the YOLOv12 model maintained consistent detection accuracy even under moderate lighting fluctuations and partial object occlusions, such as documents or equipment blocking parts of PPE. Additionally, the interface was found to be responsive and interpretable by non-technical staff. Informal feedback from archive personnel suggested that the color-coded bounding boxes and compliance classification display were helpful for quick assessments. However, challenges such as glare from reflective surfaces and the presence of non-standard PPE items were noted. These insights affirm the model’s practical potential, while also highlighting areas for enhancement in future development.

The Streamlit-based GUI was tested during real-time deployment in a hospital archive setting using a webcam feed at 720p resolution. The system achieved an average inference frame rate of 20–22 FPS on an NVIDIA RTX 3060, with an observed latency of approximately 50–70 milliseconds from frame capture to detection rendering. The interface remained responsive under continuous input, and detection updates were rendered with minimal delay. Although formal usability testing was not conducted, initial feedback from archive room personnel suggested that the layout, compliance status display, and visual indicators were easy to interpret and aligned with their workflow needs. These findings indicate that the system has promising potential for operational use, with future iterations planned for broader user testing and deployment on lower-spec hardware.

To assess the advantages of the proposed model over previous YOLO versions, a performance evaluation was conducted using a consistent dataset. This evaluation included key metrics such as precision, recall, mAP@0.5, and mAP@0.5:0.95. The comparative performance results are presented in table 5.

**Table 5.** Comparison of YOLO Model Performance with the Same Dataset

| Model | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 |
|---|---|---|---|---|
| YOLOv8 | 0.76 | 0.65 | 0.715 | 0.418 |
| YOLOv9 | 0.78 | 0.68 | 0.741 | 0.452 |
| YOLOv10 | 0.79 | 0.72 | 0.768 | 0,487 |
| YOLOv11 | 0.81 | 0.75 | 0.791 | 0.519 |
| Research Model | 0.853 | 0.836 | 0.813 | 0.570 |

Table 5 shows that the research model using YOLOv12 produces the best performance across all evaluation metrics, with a precision of 0.853, recall of 0.836, mAP@0.5 of 0.813, and mAP@0.5:0.95 reaching 0.57. Compared to the previous version, the performance improvement appears consistent in each iteration of YOLO, but a significant spike occurs in YOLOv12 thanks to the integration of the Area Attention architecture, R-ELAN, and optimization using SGD. This proves that the latest developments not only accelerate detection but also substantially improve the accuracy of object detection. This performance strengthens the validity of the YOLOv12 model as a superior approach for real-time applications in the context of occupational safety. YOLOv12 outperforms YOLOv11 with a higher mAP@0.5–0.95 score (0.570 vs. 0.519), indicating improved detection accuracy, especially for small or complex objects. The 0.051 gain highlights better spatial precision, likely due to architectural enhancements like FlashAttention and R-ELAN that improve multi-scale feature representation in cluttered indoor scenes. While the numerical improvements in mAP@0.5 and mAP@0.5–0.95 appear moderate (e.g., an increase of 0.051 over YOLOv11), these gains translate into more reliable detection of small, partially occluded, or low-contrast safety elements such as gloves and scattered cables. Such detection robustness is critical in hospital archive rooms, where visual clutter and lighting inconsistency are common. Furthermore, these gains were achieved without a significant increase in inference time, supporting the practical deployment of YOLOv12 in real-time monitoring systems. The relatively low detection accuracy for Window and Gloves is likely due to their underrepresentation in the dataset and visual similarity to the background. More effective solutions include class-specific augmentation (e.g., brightness shift, occlusion simulation) and synthetic image generation to increase data diversity. Additionally, transfer learning using pre-trained YOLOv12 weights fine-tuned on a healthcare-specific dataset can further improve performance on challenging classes. These methods will be considered in future developments to address class imbalance in a visually meaningful way. It is important to note that the reported performance metrics such as mAP@0.5 and mAP@0.5–0.95 are based on a single evaluation run and do not include standard deviation or confidence intervals. As such, the robustness of the model's performance under varying initialization or dataset splits remains to be fully validated.

While mAP@0.5 serves as a common detection metric, it may overestimate performance by accepting relatively loose bounding boxes. To address this, we also report mAP@0.5–0.95, which aggregates performance across stricter IoU thresholds (0.5 to 0.95). This provides a better proxy for localization quality. However, our study did not include a direct analysis of IoU score distributions across predictions, which would provide further insight into the tightness and spatial accuracy of bounding boxes. Future work may incorporate IoU histograms and per-class IoU statistics to evaluate and refine boundary localization performance more comprehensively.

To support real-time application claims, we measured the average inference time per image and frames per second (FPS) using an NVIDIA RTX 3060 GPU. The YOLOv12 model achieved an average inference time of approximately 45 ms per frame, or 22–25 FPS, which is sufficient for real-time detection in moderately dynamic environments such as hospital archive rooms. While testing on edge devices was not conducted in this study, published benchmarks from Ultralytics report that YOLOv12 can achieve around 6–8 FPS on Jetson Nano and 10–12 FPS on mid-range mobile CPUs with INT8 optimization. These results suggest that the model could be deployed in lightweight embedded systems with acceptable latency, though further validation is required. Future work will focus on full implementation and optimization for such platforms.

## 4. Conclusion

This study successfully developed a YOLOv12-based occupational safety detection model that is optimized to identify important safety objects such as APAR, evacuation signs, windows, and PPE in the form of masks, gloves, and shoes in the archive room environment. The results of model training showed superior performance compared to the previous YOLO version, with the final evaluation achievement being a precision of 0.853, a recall of 0.836, mAP@0.5 of 0.813, and mAP@0.5–0.95 of 0.570. Visualization through the Streamlit interface shows that the system is able to provide real-time prediction output with informative and structured room compliance level classification. The advantages of YOLOv12 in terms of architectural efficiency, application of attention areas, and optimization with SGD techniques have been proven to significantly increase the accuracy and speed of system inference.

However, limitations are still found in terms of detection precision in minor classes such as windows and gloves, which show relatively lower prediction values due to the imbalance in the amount of data and visual similarity between classes. Therefore, further research can be focused on improving the representation of minor class data, using data balancing techniques such as focal loss or oversampling, and exploring the combination of YOLOv12 with segmentation or attention fusion architectures to support more complex spatial classification. In addition, expanding the implementation of the system into a real-time hospital environment and integrating with IoT sensors are also relevant development potentials to support digital transformation in OSH monitoring, using Jetson Nano or Raspberry Pi for real-time inference in hospital rooms, integrating camera data and IoT sensors. To improve the performance of underrepresented classes, future work should explore targeted augmentation techniques, selective oversampling, and class-weighted loss functions such as focal loss to reduce the impact of class imbalance. Future work should compare optimizers like Adam and AdamW to validate the choice of SGD and identify the most effective strategy for safety object detection in constrained settings. Perform temporal modeling to detect persistent risks such as blocked exits and repeated non-use of PPE. It will also further improve label consistency for similar objects using consensus annotation and model-assisted relabeling.

## 5. Declarations

### 5.1. Author Contributions

Conceptualization: D.J., H.O.; Methodology: D.J., M.; Software: D.J.; Validation: H.O., Y.I.; Formal Analysis: D.J.; Investigation: D.J.; Resources: H.O., M., Y.I.; Data Curation: D.J.; Writing – Original Draft Preparation: D.J.; Writing – Review and Editing: H.O., M., Y.I.; Visualization: D.J.; All authors have read and agreed to the published version of the manuscript.

### 5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### 5.3. Funding

This research is funded by the Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology with the Regular Fundamental Research Scheme in 2025, in accordance with Research Contract Number: 010/LL17/DT.05.00/PL/2025.

### 5.4. Institutional Review Board Statement

Not applicable.

### 5.5. Informed Consent Statement

Not applicable.

### 5.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] H. A. Ekrami, B. F. Dehaghi, S. Ghanbari, N. J. Haghighifard, and M. J. Mohammadi, “Health risk assessment and occupational safety at hospitals in Southwest of Iran,” *Clin. Epidemiol. Glob. Heal.*, vol. 26, no. March, pp. 1-11, 2024.

[2] F. Kartika and S. L. R. Nasution, “Assessing the implementation of occupational safety and health management system in the hospital,” *J. Prima Med. Sains*, vol. 6, no. 2, pp. 244–250, 2024, doi: 10.34012/jpms.v6i2.6364.

[3] Z. Jiao, P. Hu, H. Xu, and Q. Wang, “Machine Learning and Deep Learning in Chemical Health and Safety: A Systematic Review of Techniques and Applications,” *ACS Chem. Heal. Saf.*, vol. 27, no. 6, pp. 316–334, Nov. 2020, doi: 10.1021/acs.chas.0c00075.

[4] B. Saputra, A. Utami, E. Edriyansyah, and Y. Irawan, “Expert System For Diagnosing Diseases in Toddlers Using The Certainty Factor Method,” *J. Appl. Eng. Technol. Sci.*, vol. 4, no. 1, pp. 32–41, 2022, doi: 10.37385/jaets.v4i1.916.

[5] M. Nain, S. Sharma, and S. Chaurasia, “Safety and Compliance Management System Using Computer Vision and Deep Learning,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1099, no. 1, pp. 1-10, Mar. 2021, doi: 10.1088/1757-899X/1099/1/012013.

[6] T. Zhou, Y. Niu, H. Lu, C. Peng, Y. Guo, and H. Zhou, “Vision transformer: To discover the ‘four secrets’ of image patches,” *Inf. Fusion*, vol. 105, no. May, pp. 1-12, 2024

[7] Y. Li, Z. Hu, Y. Zhang, J. Liu, W. Tu, and H. Yu, “DDEYOLOv9: Network for Detecting and Counting Abnormal Fish Behaviors in Complex Water Environments,” *Fishes*, vol. 9, no. 6, pp. 1–21, 2024, doi: 10.3390/fishes9060242.

[8] Lui MH, Liu H, Tang Z, Yuan H, Williams D, Lee D, Wong KC, Wang Z, “An Adaptive YOLO11 Framework for the Localisation, Tracking, and Imaging of Small Aerial Targets Using a Pan–Tilt–Zoom Camera Network,” *Eng*, vol. 5, no. 4, pp. 3488–3516, 2024, doi: 10.3390/eng5040182.

[9] A. Elesawy, E. Mohammed Abdelkader, and H. Osman, “A Detailed Comparative Analysis of You Only Look Once-Based Architectures for the Detection of Personal Protective Equipment on Construction Sites,” *Eng*, vol. 5, no. 1, pp. 347–366, 2024, doi: 10.3390/eng5010019.

[10] J.-H. Lo, L.-K. Lin, and C.-C. Hung, “Real-Time Personal Protective Equipment Compliance Detection Based on Deep Learning Algorithm,” *Sustainability*, vol. 15, no. 1, pp. 1-10, 2023, doi: 10.3390/su15010391.

[11] Ahmed MIB, Saraireh L, Rahman A, Al-Qarawi S, Mhran A, Al-Jalaoud J, Al-Mudaifer D, Al-Haidar F, AlKhulaifi D, Youldash M, “Personal Protective Equipment Detection: A Deep-Learning-Based Sustainable Approach,” *Sustainability*, vol. 15, no. 18, pp. 1-12, 2023, doi: 10.3390/su151813990.

[12] M. I. Al-Khiami and M. M. Elhadad, “Enhancing Construction Site Safety Using Ai: the Development of a Custom Yolov8 Model for Ppe Compliance Detection,” *Proc. Eur. Conf. Comput. Constr.*, vol. 2024, no. July, pp. 577–584, 2024, doi: 10.35490/EC3.2024.307.

[13] J. Chen, P. Shi, M. Xu, Y. Xin, X. Fan, and J. Zhang, “WCANet: An Efficient and Lightweight Weight Coordinated Adaptive Detection Network for UAV Inspection of Transmission Line Accessories,” *Drones*, vol. 9, no. 4, pp. 1-12, 2025, doi: 10.3390/drones9040318.

[14] J. Song, D. Kim, E. Jeong, and J. Park, “Determination of Optimal Dataset Characteristics for Improving YOLO Performance in Agricultural Object Detection,” *Agriculture*, vol. 15, no. 7, pp. 1-10, 2025, doi: 10.3390/agriculture15070731.

[15] A. Ghahremani, S. D. Adams, M. Norton, S. Y. Khoo, and A. Z. Kouzani, “Detecting Defects in Solar Panels Using the YOLO v10 and v11 Algorithms,” *Electron.*, vol. 14, no. 2, pp. 1-12, 2025, doi: 10.3390/electronics14020344.

[16] A. Febriani, R. Wahyuni, Y. Irawan, and R. Melyanti, “Improved Hybrid Machine and Deep Learning Model for Optimization of Smart Egg Incubator,” *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1052–1068, 2024, doi: 10.47738/jads.v5i3.304.

[17] Y. Lee and H. Kim, “Comparative Analysis of YOLO Series ( from V1 to V11 ) and Their Application in Computer Vision,” *J. Semicond. Disp. Technol.*, vol. 23, no. 4, pp. 190–198, 2024, doi: 10.3390/s25072270

[18] M. Y. Turali, A. T. Koc, and S. S. Kozat, “Optimal stochastic gradient descent algorithm for filtering,” *Digit. Signal Process.*, vol. 155, no. December, pp. 1-10, 2024.

[19] S. V Kozyrev, I. A. Lopatin, and A. N. Pechen, “Control of Overfitting with Physics,” *Entropy*, vol. 26, no. 12, pp. 1-12, 2024, doi: 10.3390/e26121090.

[20] Y. Tian, Y. Zhang, and H. Zhang, “Recent Advances in Stochastic Gradient Descent in Deep Learning,” *Mathematics*, vol. 11, no. 3, pp. 1–23, 2023, doi: 10.3390/math11030682.

[21] Buga, R., Buzea, C. G., Agop, M., Ochiuz, L., Vasincu, “Streamlit Application and Deep Learning Model for Brain Metastasis Monitoring After Gamma Knife Treatment,” *Biomedicines*, vol. 13, no. 2, pp. 1-13, 2025, doi: 10.3390/biomedicines13020423.

[22] C. Catargiu, N. Cleju, and I. B. Ciocoiu, “A Comparative Performance Evaluation of YOLO-Type Detectors on a New Open Fire and Smoke Dataset,” *Sensors*, vol. 24, no. 17, pp. 1-12, 2024, doi: 10.3390/s24175597.

[23] E. Hassan and H. Ghadiri, “Advancing brain tumor classification: A robust framework using EfficientNetV2 transfer learning and statistical analysis,” *Comput. Biol. Med.*, vol. 185, no. February, pp. 1-13, 2025.

[24] D. Rastogi, P. Johri, and V. Tiwari, “Augmentation based detection model for brain tumor using VGG 19,” *Int. J. Comput. Digit. Syst.*, vol. 13, no. 1, pp. 1227–1237, 2023, doi: 10.12785/ijcds/1301100.

[25] D. Ribeiro, D. Tavares, E. Tiradentes, F. Santos, and D. Rodriguez, “Performance Evaluation of YOLOv11 and YOLOv12 Deep Learning Architectures for Automated Detection and Classification of Immature Macauba (Acrocomia aculeata) Fruits,” *Agriculture*, vol. 15, no. 15, pp. 1-15, 2025, doi: 10.3390/agriculture15151571.

[26] Kumar, Y., Garg, P., Moudgil, M.R, “Enhancing parasitic organism detection in microscopy images through deep learning and fine-tuned optimizer,” *Sci. Rep.*, vol. 14, no. 1, pp. 1-13, 2024, doi: 10.1038/s41598-024-56323-8.

[27] D. Setiawan, R. N. Putri, I. Fitri, A. N. Hidayanto, Y. Irawan, and N. Hohashi, “Improved Deep Learning Model for Prediction of Dermatitis in Infants,” *J. Appl. Data Sci.*, vol. 6, no. 2, pp. 871–884, 2025, doi: 10.47738/jads.v6i2.542.

[28] Anam, M. K., L. L. Van FC, H. Hamdani, R. Rahmaddeni, J. Junadhi, M. B. Firdaus, I. Syahputra, and Y. Irawan, “Sara Detection on Social Media Using Deep Learning Algorithm Development ,” *J. Appl. Eng. Technol. Sci.*, vol. 6, no. 1 SE-Articles, pp. 225–237, Dec. 2024, doi: 10.37385/jaets.v6i1.5390.

[29] A. Ali, W. Ou, and S. Kanwal, “DCTNets: Deep crowd transfer networks for an approximate crowd counting,” *Cogn. Robot.*, vol. 2, no. April, pp. 96–111, 2022, doi: 10.1016/j.cogr.2022.03.004.

[30] F. S. Nahm, “Receiver operating characteristic curve: overview and practical use for clinicians,” *kja*, vol. 75, no. 1, pp. 25–36, Jan. 2022, doi: 10.4097/kja.21209.

[31] A. Lubis, Y. Irawan, Junadhi, and S. Defit, “Leveraging K-Nearest Neighbors with SMOTE and Boosting Techniques for Data Imbalance and Accuracy Improvement,” *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 1625–1638, 2024, doi: 10.47738/jads.v5i4.343.

[32] H. Fonda, Y. Irawan, R. Melyanti, R. Wahyuni, and A. Muhaimin, “A Comprehensive Stacking Ensemble Approach for Stress Level Classification in Higher Education,” *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 1701–1714, 2024, doi: 10.47738/jads.v5i4.388.

[33] Herianto, B. Kurniawan, Z. H. Hartomi, Y. Irawan, and M. K. Anam, “Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction,” *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1272–1285, 2024, doi: 10.47738/jads.v5i3.316.