Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 6, December 2025, Page. 5699-5790
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5419

Interpretable Hybrid YOLOv8s-GWO Framework for Bounding-Box Viral Pneumonia Detection on Kaggle Chest X-ray Images

Azmi Jalaluddin Amron[1], Cinantya Paramita*[2], Petar Šolić[3], Supratiknyo[4]

[1]Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[2]Dinus Research Group for AI in Medical Science (DREAMS), Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[3] Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Split, Croatia
[4]SMK Sunan Kalijaga Demak, Demak, Indonesia

Email: [2]cinantya.paramita@dsn.dinus.ac.id

Received : Oct 31, 2025; Revised : Nov 30, 2025; Accepted : Dec 16, 2025; Published : Dec 23, 2025

Abstract

Viral pneumonia continues to impose a substantial global health burden, making rapid and reliable radiographic detection essential for early clinical management. This study proposes a hybrid framework integrating the YOLOv8s detection model with the Grey Wolf Optimizer (GWO) to enhance hyperparameter tuning for Viral Pneumonia identification in chest X-ray images. A curated set of Normal and Viral Pneumonia samples was manually annotated and preprocessed before training. The optimization process involved multi-stage refinement of learning rate, momentum, weight decay, and loss-gain parameters to improve convergence stability and detection accuracy. The optimized YOLOv8s + GWO model demonstrated notable performance gains, achieving 0.965 recall, 0.983 mAP@50, and 0.827 mAP@50–95 on internal evaluations. External testing further validated its robustness, delivering 98.80% accuracy, 99.48% specificity, and 97.46% sensitivity. These results highlight not only enhanced clinical diagnostic reliability but also contributions to Informatics and Computer Science, demonstrating the effectiveness of metaheuristic-guided optimization in improving deep-learning model performance, generalization, and computational efficiency for AI-driven image detection tasks.

Keywords : *Grey Wolf Optimizer, Hyperparameter Optimization, Medical Image Detection, Viral Pneumonia, YOLOv8s, Chest X-ray.*

1. INTRODUCTION

Pneumonia is a severe respiratory infection characterized by alveolar inflammation, leading to fluid accumulation and impaired gas exchange in the lungs [1], [2]. It remains a leading cause of infectious disease mortality worldwide, accounting for over 2.5 million deaths annually, including approximately 672,000 children under five years of age [1]. Regional epidemiological studies in East Asia indicate that lower respiratory tract infections continue to represent a major cause of hospital admissions, particularly among pediatric and elderly populations [2], [3]. Multiplex RT-PCR analyses further demonstrate that Viral pathogens contribute significantly to seasonal respiratory infections [3], [5]. This epidemiological burden underscores the urgent clinical need for rapid and accurate differentiation between Viral and Bacterial Pneumonia, a task that remains challenging in settings with limited radiological expertise.

Radiographically, Viral Pneumonia commonly manifests as diffuse, bilateral interstitial infiltrates or ground-glass opacities, whereas bacterial pneumonia often presents as localized consolidations [4]. These subtle imaging differences can result in considerable diagnostic variability. Integrating artificial intelligence (AI) into radiology workflows offers a potential solution for automated feature extraction and rapid interpretation of chest imaging [5], [6], [19]. The YOLO (You Only Look Once) family of deep learning models has emerged as a leading architecture due to its real-time detection capabilities, high spatial precision, and proven performance across a range of medical imaging applications [6], [7], [10], [11], [15], [16], [18], [27]. Recent research has expanded YOLOv8 applications to multi-organ detection in chest X-rays [46], real-time detection of lung diseases [12], and enhanced A-line and B-line detection in lung ultrasound [39], [40], highlighting the trend toward ensemble and multi-modal AI strategies in medical imaging.

Despite these advances, YOLO-based models often experience performance degradation when confronted with heterogeneous image quality, domain shifts, or class imbalance [6], [12], [15]. In radiology, AI-powered object detection has attracted significant attention for its potential to improve diagnostic workflows [19]. Beyond medical imaging, YOLO architectures have been successfully applied to a variety of tasks, including skin lesion detection [29-31], cataract detection [33-34], autonomous driving [38], traffic monitoring [35], and herbal product identification [37], highlighting their adaptability across diverse computer vision applications. However, manual hyperparameter tuning can result in suboptimal convergence, unstable sensitivity, and limited generalization. To address this, metaheuristic optimization algorithms particularly the Grey Wolf Optimizer (GWO) have shown promise in enhancing hyperparameter selection for deep learning models [13], [14], [17], [20], [43], [44]. Leveraging GWO’s global search capabilities can stabilize training and improve diagnostic accuracy across healthcare tasks [12], [14], [17], [20]. Nevertheless, its application to multi-modal datasets, such as the integration of chest X-ray and lung ultrasound imaging, remains largely unexplored, restricting the potential for achieving higher accuracy and better generalization [6], [12], [39], [40].

Table 1 presents a summary of the publicly available pneumonia datasets used in this study, providing a clear overview of the number of images and class distribution for both training and testing.

Table 1. Composition of Pneumonia Training and Testing Datasets

Dataset	Label	Number of Images
Training [8]	Normal	1,583
Training [8]	Viral Pneumonia	1,493
Testing [9]	Normal + Viral Pneumonia	4,926

Figures 1 and 2 illustrate representative radiographic differences between normal lungs and Viral Pneumonia. Normal chest X-rays display clear lung fields with well-defined anatomical structures, while Viral Pneumonia images exhibit diffuse or patchy opacities that obscure lung anatomy. These visual cues are essential for training YOLOv8s to detect and highlight abnormal regions in a clinically interpretable manner.

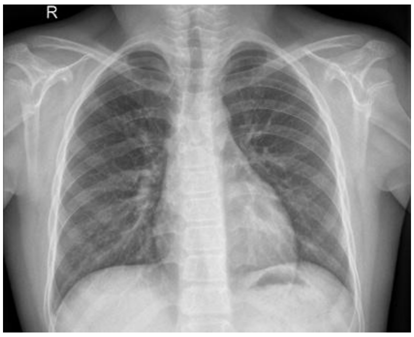

Figure 1. Normal Lungs

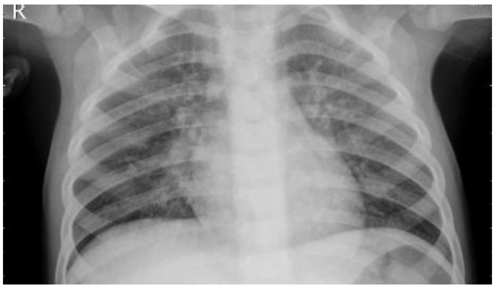

Figure 2. Viral Pneumonia Lungs

Conventional YOLOv8s training on heterogeneous chest X-ray datasets achieved high mAP@50 scores (~97%) but exhibited lower overall performance across multiple thresholds, with mAP@50-95 around 81-82% [6], [12]. Manual hyperparameter tuning provided limited improvement, often yielding marginal gains in mAP and recall. Integrating the Grey Wolf Optimizer (GWO) demonstrated the potential to enhance model performance, improving mAP@50-95, accuracy, and specificity, while maintaining comparable recall and sensitivity. In this study, the YOLOv8s + GWO framework achieved an accuracy of 98.8%, specificity of 99.48%, recall of 95.55%, and mAP@50-95 of 82.1%, indicating reliable detection of Viral Pneumonia across diverse datasets. Model performance was comprehensively assessed using Precision, Recall, Accuracy, Specificity, and mAP, ensuring robust evaluation and minimizing false negatives, which is critical for clinical decision-making.

This study proposes a novel hybrid YOLOv8s + GWO framework for Viral pneumonia detection. Unlike prior approaches that rely solely on YOLOv8s or manual tuning, the proposed method leverages GWO to simultaneously optimize multiple hyperparameters, including learning rate, momentum, and weight decay. By addressing multi-modal limitations, enhancing training stability, and improving generalization, the framework is designed to deliver Viral clinically meaningful predictions, supporting rapid diagnosis and optimized workflows in resource-constrained healthcare settings. The integration of AI-driven informatics engineering demonstrates the potential to enhance early detection, improve clinical decision-making, and extend deep learning applicability across heterogeneous imaging modalities [6], [12], [14], [18]-[28], [39], [40], [46]-[49].

2. METHOD

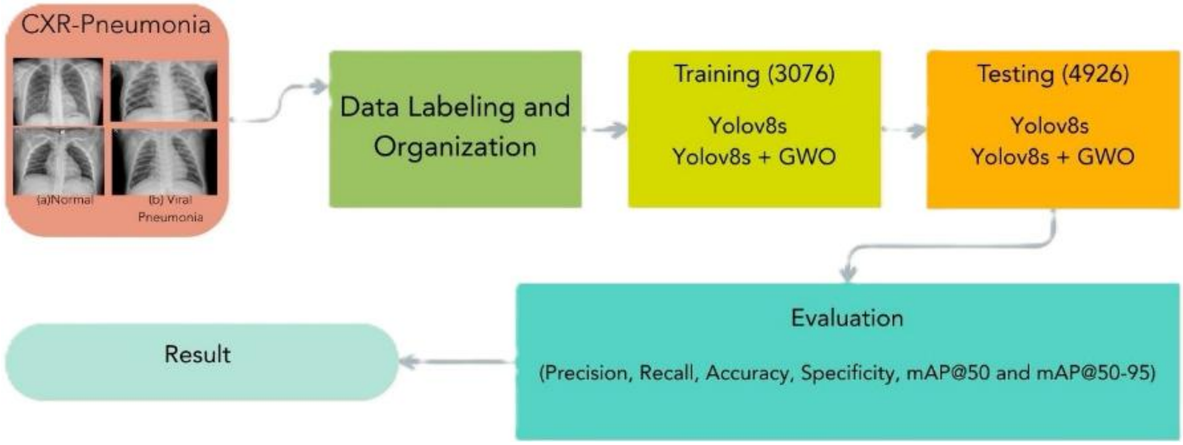

Figure 3. End-to-End YOLOv8s-GWO Pipeline with 3-Phase Optimization

Figure 3 illustrates the complete methodological pipeline of the proposed YOLOv8s–GWO framework for Viral Pneumonia detection from chest X-ray images. The workflow begins with dataset preparation and annotation, followed by preprocessing and augmentation, baseline YOLOv8s training, multi-stage Grey Wolf Optimizer (GWO) hyperparameter tuning, final retraining using the optimized configuration, and model evaluation. This structured design ensures reproducibility, systematic hyperparameter exploration, and robust performance assessment.

2.1. Dataset Preparation, Annotation, and Ethical Considerations

Two publicly available chest X-ray datasets were initially considered, namely the Chest X-ray Pneumonia dataset [8] and the Three Kinds of Pneumonia dataset [9]. However, for model training, only dataset [8] was utilized, as it contains clearly separated image groups for Normal, Viral Pneumonia, and Bacterial Pneumonia. Following prior studies focusing on Viral Pneumonia detection [6], [7], [10], [12], all Bacterial Pneumonia samples were excluded, resulting in a total of 3,076 images, comprising 1,583 Normal and 1,493 Viral Pneumonia cases.

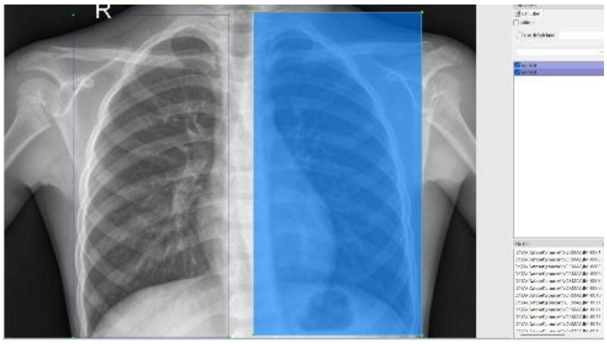

Figure 4. Annotation interface of *LabelImg* showing a chest X-ray image labeled as 'Normal'

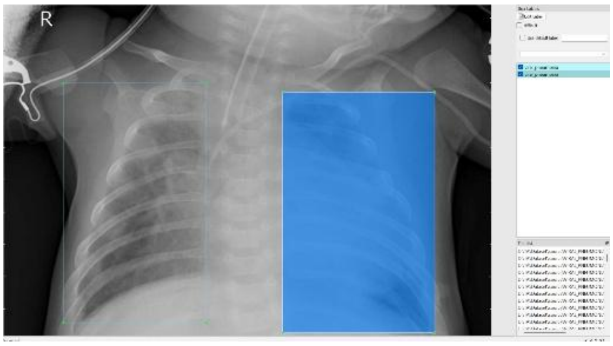

Figure 5. Annotation interface of *LabelImg* showing a chest X-ray image labeled as 'Viral Pneumonia

All images were manually annotated using *LabelImg* v1.8.6 to generate YOLO-format text files containing class identifiers and normalized bounding-box coordinates. Figures 4 and 5 illustrate representative annotated samples for both classes. After filtering out corrupted, duplicated, and low quality images, the remaining dataset was divided into 2,139 images for training, 611 images for validation, and 307 images for testing, preserving class balance and reducing distributional bias during model development.

For model evaluation, the Three Kinds of Pneumonia dataset [9] was utilized as the test set. Only the Normal and Viral Pneumonia classes were selected from this dataset, resulting in 3,270 Normal images and 1,656 Viral Pneumonia images. These images were combined into a single folder named *test_xray* to serve as an independent test set for assessing the model's performance after training.

All datasets used in this study were publicly available on Kaggle and fully anonymized. The research adhered to GDPR regulations. Since all data were publicly accessible and de-identified, the institutional review board (IRB) confirmed that formal ethical approval was not required for this study.

2.2. Preprocessing and Augmentation Strategy

All images were resized to 640 × 640 pixels and normalized according to the *Ultralytics* YOLOv8 preprocessing pipeline. To improve generalization and reduce overfitting, an extensive augmentation strategy was applied using *Albumentations* 2.0.8. This included random horizontal flipping, affine transformations, variations in brightness and contrast, slight rotations, and other regularization techniques supported by the YOLOv8 pipeline. These augmentations simulate realistic radiographic variations commonly encountered in clinical practice and help the model avoid over-reliance on specific image regions, improving robustness to unseen data.

2.3. Baseline YOLOv8s Architecture and Initial Training

The YOLOv8s architecture was chosen for its computational efficiency and strong performance in medical image detection tasks [10], [11], [15], [16], [27]. It features an end-to-end design that seamlessly integrates convolutional feature extraction, bounding box regression, and object classification within a single unified network. A baseline YOLOv8s model was trained for 100 epochs using the default hyperparameters provided in *Ultralytics* YOLOv8 v8.3.206, serving as the reference against which the GWO-optimized configuration was compared. Previous studies have highlighted YOLOv8's effectiveness in detecting small objects [15], tuberculosis [16], and pulmonary nodules [42], while metaheuristic methods such as GWO have been successfully applied for image optimization and super-resolution [43, 44]. These findings inspired the integration of YOLOv8 with GWO to enhance pneumonia detection performance.

2.4. Grey Wolf Optimizer (GWO) for Hyperparameter Tuning

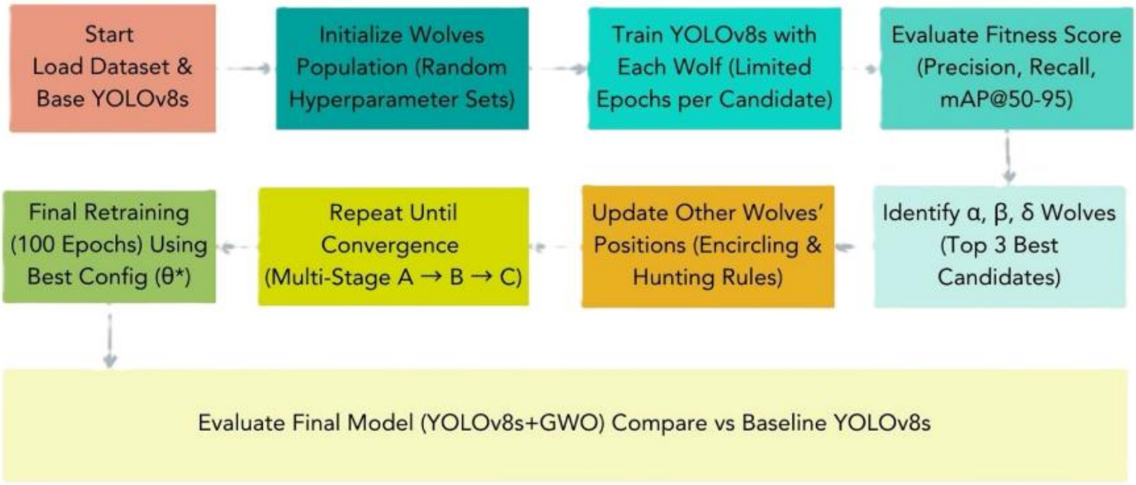

Figure 6. GWO Encircling-Hunting Phases for YOLOv8s Hyperparameter Search Space

Figure 6 illustrates the overall hyperparameter optimization process powered by the Grey Wolf Optimizer. GWO is inspired by the social hierarchy and cooperative hunting strategies of grey wolves [12], [13], and was employed to explore a six dimensional hyperparameter search space for YOLOv8s. Each wolf in the population encodes a candidate configuration represented as

$$\theta_i = \{lr_0, momentum, weigh\_decay, box, cls, optimizer\} \tag{1}$$

where the optimizer parameter is sampled from the discrete set {SGD, Adam, AdamW}. The performance of each candidate configuration is evaluated using the objective function

$$f(\emptyset_i) = 0.1P + 0.1R + 0.7mAP_{50-95} + 0.05mAP_{50-95}^{Normal} + 0.05mAP_{50-95}^{Viral} \tag{2}$$

which prioritizes overall localization accuracy while preserving balanced performance across the Normal and Viral classes [9], [11].

The encircling behavior of grey wolves is modeled as:

$$\overrightarrow{D_\alpha} = \left|\overrightarrow{C_1} \cdot \overrightarrow{X_\alpha} - \vec{X}(t)\right|, \qquad \overrightarrow{X_1} = \overrightarrow{X_\alpha} - \overrightarrow{A_1} \cdot \overrightarrow{D_\alpha} \tag{3}$$

$$\overrightarrow{D_\beta} = \left|\overrightarrow{C_2} \cdot \overrightarrow{X_\beta} - \vec{X}(t)\right|, \qquad \overrightarrow{X_2} = \overrightarrow{X_\beta} - \overrightarrow{A_2} \cdot \overrightarrow{D_\beta} \tag{4}$$

$$\overrightarrow{D_\delta} = \left|\overrightarrow{C_3} \cdot \overrightarrow{X_\delta} - \vec{X}(t)\right|, \qquad \overrightarrow{X_3} = \overrightarrow{X_\delta} - \overrightarrow{A_3} \cdot \overrightarrow{D_\delta} \tag{5}$$

The position of each wolf is updated by averaging the influence of the three best solutions:

$$\overrightarrow{X(t+1)} = \frac{\overrightarrow{X_1}+\overrightarrow{X_2}+\overrightarrow{X_3}}{3} \tag{6}$$

The coefficient vectors $\vec{A}$ and $\vec{C}$ are computed as:

$$\vec{A} = 2a\vec{r}_1 - a, \qquad \vec{C} = 2\vec{r}_2 \tag{7}$$

Where α decreases linearly from 2 to 0 over iterations and $\overrightarrow{r_1}, \overrightarrow{r_2}$ are random vectors in [0,1].

A structured three phase optimization strategy was adopted. Stage A performs global exploration with relatively broad variations in hyperparameters. Stage B emphasizes regional refinement by narrowing the search region around high performing candidates. Stage C conducts fine tuning with increased evaluation fidelity. Each stage consists of a quick evaluation phase followed by a long evaluation phase to ensure balanced assessment of both promising and potentially overlooked configurations. Table 2 summarizes the characteristics of the three stages.

Table 2. Three-stages GWO Optimization Strategy

Stage	Quick Eval	Long Eval	Purpose
A	20 epochs	20 epochs	Global exploration
B	40 epochs	40 epochs	Regional refinement
C	60 epochs	60 epochs	Local fine-tuning

Table 3. Stepwise GWO-based Hyperparameter Optimization Procedure for YOLOv8s

Step	Description
Input:	Dataset configuration (*dataset.yaml*), YOLOv8s base model (yolov8s.pt), hyperparameter ranges ($\boldsymbol{R}$), number of wolves ($\boldsymbol{N}$), and maximum iterations ($\boldsymbol{T}$).
Output:	Optimized hyperparameter configuration ($\boldsymbol{\theta}^*$) for YOLOv8s + GWO.
1. Initialize Population	Randomly generate ($\boldsymbol{N}$) wolves, each representing a parameter set $\boldsymbol{\theta_i} = \{\boldsymbol{lr_0}, \mathbf{momentum}, \mathbf{weight\_decay}, \mathbf{box}, \mathbf{cls}, \mathbf{optimizer}\}$ sampled from ($\boldsymbol{R}$).
2. Evaluate Initial Fitness	Train YOLOv8s using each $\boldsymbol{\theta_i}$ for limited epochs ($\boldsymbol{E}$) and compute the fitness score $\boldsymbol{f}(\boldsymbol{\theta_i})$.
3. Identify Elite Wolves	Select top three wolves as α (best), β (second), and δ (third).
4. Update Positions (Iterative Optimization)	For each non-elite wolf, update its position using GWO's encircling mechanism and randomly adjust optimizer type among $\{\boldsymbol{SGD}, \boldsymbol{Adam}, \boldsymbol{AdamW}\}$.
5. Re-evaluate Fitness	Train YOLOv8s using updated $\boldsymbol{\theta_i}(\boldsymbol{t}+\mathbf{1})$, compute new fitness $\boldsymbol{f}(\boldsymbol{\theta_i}(\boldsymbol{t}+\mathbf{1}))$, and update ranking.
6. Repeat Iterations	Continue updating and evaluating until reaching ($\boldsymbol{T}$) iterations or convergence.
7. Multi-Stage Refinement	Conduct optimization in three stages: Stage A (20 epochs), Stage B (40 epochs), Stage C (60 epochs).
8. Final Retraining	Train YOLOv8s using the best configuration ($\boldsymbol{\theta}^*$) for 100 epochs to obtain the optimized YOLOv8s + GWO model.

To provide a clear and systematic overview of the multi-stage GWO optimization process, the stepwise procedure is summarized in Table 3. Each step outlines the initialization of the wolf population, evaluation of candidate hyperparameter configurations, identification of elite wolves, iterative position updates via encircling and hunting mechanisms, multi-stage refinement, and final retraining of the YOLOv8s model using the optimized hyperparameters. Referring to Table 3 allows readers to follow the optimization workflow in a structured and reproducible manner, complementing the mathematical formulations and stage descriptions provided above.

2.4.1. Ablation Study: PSO vs GWO for Hyperparameter Optimization

To further justify the choice of the Grey Wolf Optimizer as the primary hyperparameter tuning strategy, an ablation study was conducted comparing GWO with the Particle Swarm Optimization (PSO) algorithm. Both methods were applied to the same baseline YOLOv8s model and evaluated using identical dataset partitions, training configurations, and evaluation metrics. The goal of this comparison was not to provide an exhaustive benchmarking between the two metaheuristic algorithms, but rather to confirm that GWO achieves at least comparable performance while aligning with the proposed optimization strategy.

In this study, a limited number of particles and wolves were used to ensure rapid evaluation while still reflecting the search dynamics of each algorithm. PSO was configured with three particles and three iterations, while GWO employed three wolves over three iterations. Each candidate hyperparameter set consisted of the learning rate, momentum, and weight decay. For every candidate configuration, the model was evaluated using the YOLOv8 validation pipeline to obtain the mAP50 metric.

The results of the ablation study are summarized in Table 4. As observed, both PSO and GWO reached the same maximum mAP50 of 0.98077 despite variations in the specific hyperparameter combinations explored by each optimizer. The learning rates, momentum values, and weight decay coefficients differed among the best-performing candidates of each algorithm, reflecting the inherent stochastic nature of metaheuristic searches. However, these differences did not translate into performance gaps, as the model’s validation results remained effectively equivalent.

Table 4. Ablation Study Comparing PSO and GWO Performance

Optimizer	Best Learning Rate	Best Momentum	Best Weight Decay	mAP50
PSO	0.000195	0.894	0.000367	0.98077
GWO	0.000329	0.898	0.000449	0.98077

The equivalence in performance can be attributed to the fact that both PSO and GWO effectively explore the local regions of the hyperparameter search space near the baseline configuration. Given the relatively small search ranges and the already well-tuned baseline model, the marginal differences in candidate solutions did not significantly affect the network’s ability to detect Viral Pneumonia. This finding reinforces the suitability of GWO for the task, demonstrating that it can achieve competitive results while providing the structured, multi-phase optimization framework detailed in the previous sections.

2.5. Algorithmic Workflow and Reproducibility

The complete optimization routine was implemented under strict determinism. All random seeds for Python, NumPy, and *PyTorch* were fixed to 42, and non-deterministic *CuDNN* operations were disabled. Hyperparameter search ranges were progressively tightened at the end of each stage to promote convergence.

The formal pseudocode of the proposed multi-stage GWO is presented as **Algorithm 1**.

Algorithm 1. Multi-Stage GWO for Hyperparameter Optimization

```
Require: Dataset $D$; YOLOv8s model $M$; parameter ranges $R$; number of wolves $N$; stages $S = \{A, B, C\}$; evaluation epochs $(E_{quick}, E_{long})$
Ensure: Optimized hyperparameters $\theta^*$
1: Initialize population $\Theta = \{\theta_1, \theta_2, \dots, \theta_N\}$ sampled from $R$
2: Set all random seeds for deterministic execution
3: for each stage $s \in S$ do
4:     for each wolf $\theta_i \in \Theta$ do
5:         Train $M$ with $\theta_i$ for $E_{quick}$ epochs
6:         Compute fitness $f(\theta_i)$
7:     end for
8:     Sort $\Theta$ and identify $\alpha, \beta, \delta$ wolves
9:     for each non elite wolf $\theta_i$ do
10:         Update $\theta_i$ using GWO encircling and hunting equations
```

```
        Randomly resample optimizer type $\in \{SGD, Adam, AdamW\}$
    end for
    for each wolf $\theta_i \in \Theta$ do
        Train $M$ with $\theta_i$ for $E_{long}$ epochs
        Recompute fitness $f(\theta_i)$
    end for
    Sort $\Theta$ and tighten parameter ranges around the top wolves
end for
Determine $\theta^* = \arg max_{\theta_i} f(\theta_i)$
return $\theta^*$
```

2.6. Optimal Configuration and Model Training

After completing the multi-stage optimization, the final optimal hyperparameter configuration is summarized in Table 5. This configuration includes the learning rate, momentum, *weight_decay*, bounding-box regression gain, classification gain, and optimizer type. The final YOLOv8s model was retrained for 100 epochs using this optimized configuration to produce the proposed YOLOv8s + GWO model. The optimized model demonstrated improved convergence stability, precision, recall, and overall mAP, confirming the effectiveness of the metaheuristic optimization strategy in medical imaging applications [12], [13].

Table 5. Final Optimized YOLOv8s + GWO Hyperparameter Configuration

Parameter	Description	Optimal Value
optimizer	Optimizer type	AdamW
$\boldsymbol{lr_0}$	Initial learning rate	0.009621
momentum	Momentum factor	0.9260
weight_decay	L2 regularization coefficient	0.0003699
box	Box regression gain	0.08037
cls	Classification loss gain	0.22069

2.7. Experimental Setup and Evaluation Metrics

Experiments were conducted on a Lenovo laptop equipped with an NVIDIA GeForce RTX 4050 GPU (6 GB), Intel® Core™ i7-13620H CPU (2.40 GHz), and 16 GB RAM running Windows 11 Home (64-bit). The training environment utilized CUDA 12.6, *PyTorch* 2.5.1 + cu121, and *Ultralytics* YOLOv8 v8.3.206, supported by NumPy 2.2.6, OpenCV 4.12.0.88, *Albumentations* 2.0.8, and scikit-learn 1.7.2. Training the baseline and optimized models required approximately 1.09 hours per 100 epochs.

Model performance was evaluated using Precision, Recall, Accuracy, Sensitivity, Specificity, and Mean Average Precision (mAP) [1], [4], [6], [12], [14], [18], [21], [22], [23], [24]. Sensitivity was emphasized because of its clinical importance in detecting Viral Pneumonia cases, while specificity measured false positives among normal images. All evaluations were conducted using YOLOv8's built-in validation pipeline to ensure consistent and reproducible metric computation.

Vol. 6, No. 6, December 2025, Page. 5699-5790
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5419

3. RESULT

In this section, the results of the research and the experiments carried out are presented. The results include both quantitative and qualitative analyses of the YOLOv8s and YOLOv8s + GWO models for Viral Pneumonia detection from chest X-ray images.

3.1. Dataset Preparation and Annotation

The study utilized a dataset comprising 4,254 labeled chest X-ray instances, including 2,224 Normal and 2,030 Viral Pneumonia bounding-box annotations [8]. The bounding boxes were predominantly centered within the lung regions, reflecting anatomical consistency and providing sufficient variability in size and spatial location. This spatial distribution allows the models to learn robust representations of both normal and pathological features [6][12][14][21][22].

To assess the consistency of manual annotations, Cohen's Kappa score was calculated based on multiple annotators. The resulting Kappa of 0.9993 indicates almost perfect agreement, confirming the high reliability of the dataset labeling. Table 6 presents the confusion matrix derived from this inter-rater reliability analysis, illustrating that only a single disagreement occurred for the Viral Pneumonia class.

Table 6. Inter-rater annotation agreement confusion matrix

Predicted / Actual	Normal	Viral Pneumonia
Normal	1583	0
Viral Pneumonia	1	1473

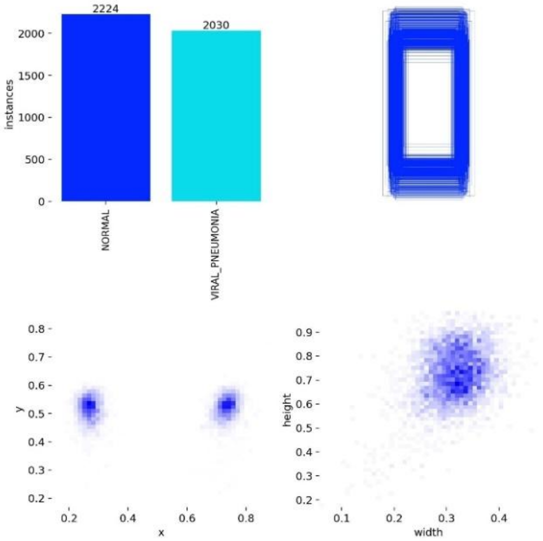

Figure 7. Class distribution of the Chest X-ray Pneumonia dataset [8]

Figure 7 illustrates the detailed class distribution and annotation characteristics for the Chest X-ray Pneumonia dataset [8]. It shows the total number of labels per class (Normal: 2,224; Viral Pneumonia: 2,030), the approximate size and spatial location of the bounding boxes within a representative image, and the corresponding class label for each annotation. This visualization highlights the comprehensive coverage of lung regions and supports the model's capacity to learn relevant features.

Figure 8 provides an overview of the complete dataset distribution, including both the training and validation images from the Chest X-ray Pneumonia dataset [8] as well as the independent test images from the Three Kinds of Pneumonia dataset [9]. The pie chart indicates the percentage composition of each subset relative to the total dataset of 8,002 images, with *train+val* Normal at approximately 19.8%, *train+val* Viral Pneumonia at 18.7%, test Normal at 40.9%, and test Viral Pneumonia at 20.7%. The accompanying bar chart presents the absolute number of images per subset (Normal [8]: 1,583; Viral Pneumonia [8]: 1,493; Normal [9]: 3,270; Viral Pneumonia [9]: 1,656), providing a clear quantitative view of dataset allocation for model training, validation, and external testing.

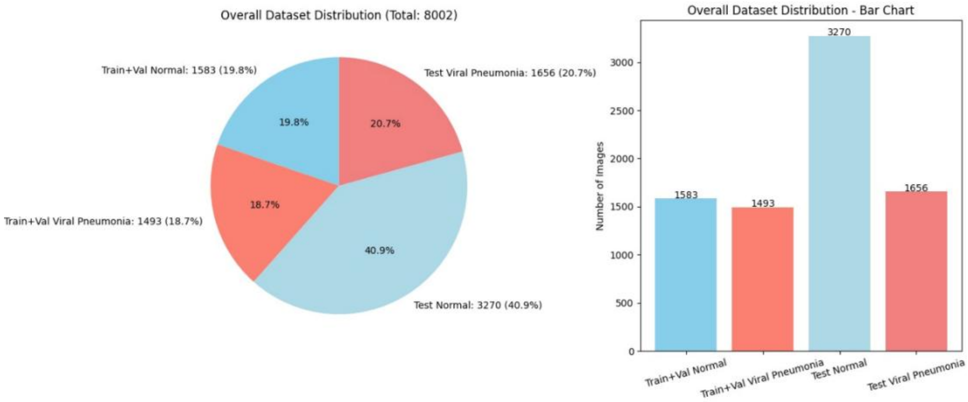

Figure 8. Distribution of training, validation, and independent test dataset across Normal and Viral Pneumonia [8][9]

For external evaluation, the Three Kinds of Pneumonia dataset [9] was employed as an independent test set. Only the Normal and Viral Pneumonia classes were selected, resulting in 3,270 Normal images and 1,656 Viral Pneumonia images, which were combined into a single folder named *test_xray* to facilitate model testing after training.

3.2. Preprocessing and Augmentation

All chest X-ray images were preprocessed following the standard YOLOv8 input pipeline, which included resizing each image to 640 × 640 pixels, normalization, and conversion into YOLO format tensors prior to model ingestion. To enhance robustness and reduce overfitting, data augmentation was applied using *Albumentations* 2.0.8. The augmentation operations included random horizontal flipping, affine transformations, brightness–contrast adjustments, and mild rotational perturbations, each selected to simulate realistic radiographic variations without distorting anatomical structures.

The same preprocessing and augmentation pipeline was applied consistently to both the baseline YOLOv8s model and the GWO optimized YOLOv8s variant. This ensures that any performance differences between the two models arise solely from hyperparameter optimization rather than from differences in data manipulation or input transformation. As a result, both models were trained under identical input conditions, guaranteeing fairness and comparability across all subsequent analyses.

3.3. Baseline YOLOv8s Training

The baseline YOLOv8s model was trained for 100 epochs using the default *Ultralytics* YOLO pipeline. Training and validation were conducted under identical hardware conditions (NVIDIA GeForce RTX 4050 GPU, Intel Core i7 processor, and 16 GB RAM). The YOLOv8s architecture used in this study contains a 72 layers backbone with approximately 11.13 million parameters, consistent with prior object detection studies in medical imaging. Both the baseline YOLOv8s and the optimized YOLOv8s + GWO models were trained using the same dataset split, preprocessing steps, and augmentation schemes to ensure fair comparison.

The training procedure followed the standard YOLOv8 optimization flow, including adaptive learning-rate scheduling, momentum-based gradient updates, and multi-scale loss computation consisting of bounding-box loss, classification loss, and objectness probability loss. Validation was conducted at every epoch to monitor convergence and prevent overfitting.

3.3.1. Performance Evaluation Metrics

Model performance was assessed using standard object detection metrics including Precision (P), Recall (R), Accuracy, Specificity, Sensitivity, Average Precision (AP), and Mean Average Precision (mAP) at IoU thresholds of mAP@50 and mAP@50-95. Sensitivity was emphasized due to its clinical importance for ensuring pneumonia cases are not missed.

Formulas for the evaluation metrics are as follows:

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

$$Specificity = \frac{TN}{TP+FP} \tag{11}$$

$$AP_i = \int_0^1 p_i(r)dr \tag{12}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{13}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

The combined summary metrics for YOLOv8s and YOLOv8s + GWO are presented in Table 7, showing overall improvements after hyperparameter optimization.

Table 7. Summary Metrics for YOLOv8s and YOLOv8s + GWO

Metric	YOLOv8s	YOLOv8s + GWO
Precision	0.95038	0.95836
Recall	0.95453	0.95553
mAP@50	0.97192	0.97727
mAP@50–95	0.81536	0.82091
Accuracy	0.95920	0.98800
Specificity	0.94530	0.99480
Sensitivity	0.98670	0.97460

This table demonstrates that the optimized model outperforms the baseline in most metrics, particularly in accuracy, specificity, and overall detection precision.

3.3.2. Training and Validation Performance

Training and validation results for both models are presented in Table 8, showing the behavior of each model during supervised optimization. YOLOv8s achieved strong baseline performance, while YOLOv8s + GWO achieved improved recall and mAP scores, indicating better generalization on unseen data.

Table 8. Training and validation performance of YOLOv8s and YOLOv8s + GWO

Model	Precision (P)	Recall (R)	mAP@50	mAP@50-95	Training Time (hours)
YOLOv8s	0.962	0.956	0.980	0.826	1.087
YOLOv8s + GWO	0.946	0.965	0.983	0.827	1.095

These results indicate that although YOLOv8s + GWO obtains slightly lower precision during training, its higher recall and mAP values suggest more stable and generalized learning behavior.

3.3.3. Stability Metrics for YOLOv8s and YOLOv8s + GWO

To evaluate training consistency, both models were executed across multiple runs, and the resulting standard deviations of key performance metrics were calculated. Lower standard deviation values indicate greater stability across training runs, while higher values indicate greater variability. The stability metrics for the two models are summarized in Table 9.

Table 9. Stability Metrics for YOLOv8s and YOLOv8s + GWO

Metric	YOLOv8s std	YOLOv8s + GWO std
Precision	0.05346	0.13074
Recall	0.06101	0.10906
mAP@50	0.04116	0.13261
mAP@50–95	0.05964	0.14051

The results indicate that the baseline YOLOv8s model exhibits higher training stability, as reflected by its consistently lower standard deviations across all metrics. In contrast, the YOLOv8s + GWO model, although achieving higher peak performance, shows increased variability between runs. This behavior is expected, given that the Grey Wolf Optimizer introduces a stronger exploratory component during hyperparameter search, leading to greater fluctuations across optimization trials.

From a practical standpoint, these findings imply that GWO prioritizes performance improvement at the possible expense of run-to-run consistency. Such behavior is typical for population-based metaheuristic optimizers, particularly when the search space is large or contains multiple local optima. Nonetheless, the improved performance obtained with GWO suggests that the trade-off between slightly increased variability and higher accuracy remains acceptable, especially for applications where maximizing detection performance is more critical than ensuring deterministic training behavior.

3.3.4. Histogram Distribution of Evaluation Metrics

To visualize how the evaluation metrics distribute across multiple runs, a combined histogram of Precision, Recall, mAP@50, and mAP@50-95 was generated for both YOLOv8s and YOLOv8s + GWO. This visualization helps highlight the consistency of model performance and the degree of improvement contributed by the GWO optimization.

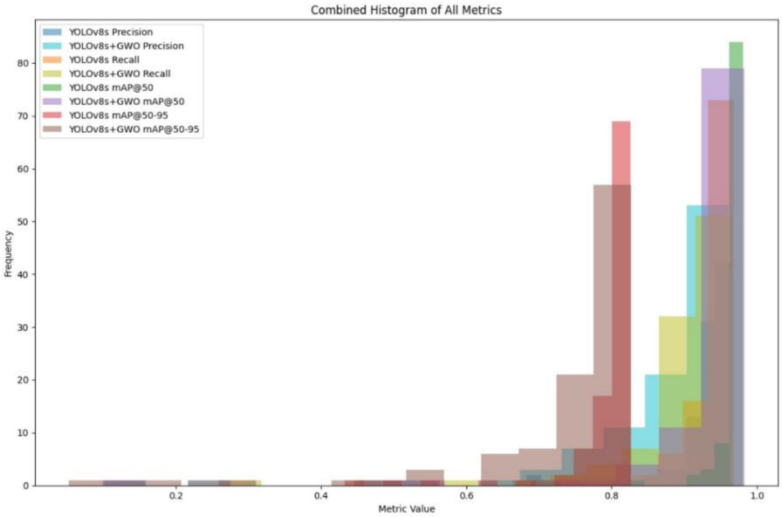

Figure 9. Combined histogram of Precision, Recall, mAP@50, and mAP@50-95 for YOLOv8s and YOLOv8s + GWO

Figure 9 illustrates that YOLOv8s + GWO exhibits a slight rightward shift in all performance metrics, indicating improved predictive quality. Although the variance appears marginally wider due to exploration during hyperparameter tuning, the optimized model maintains a generally higher and more stable distribution across key detection metrics.

3.3.5. Training Curve Analysis

Training curve plots were generated to evaluate loss functions and recall behavior across epochs. These include box loss convergence, classification or objectness probability loss behavior, bounding-box recall, and precision-recall evolution.

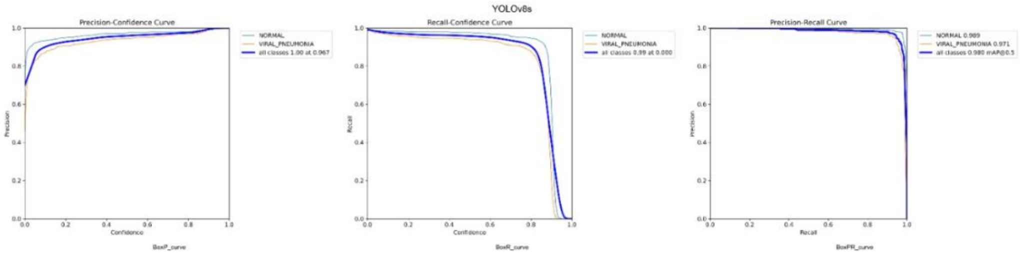

Figure 10. Training curves of YOLOv8s showing Box Probability Loss, Bounding-Box Recall, and Precision-Recall

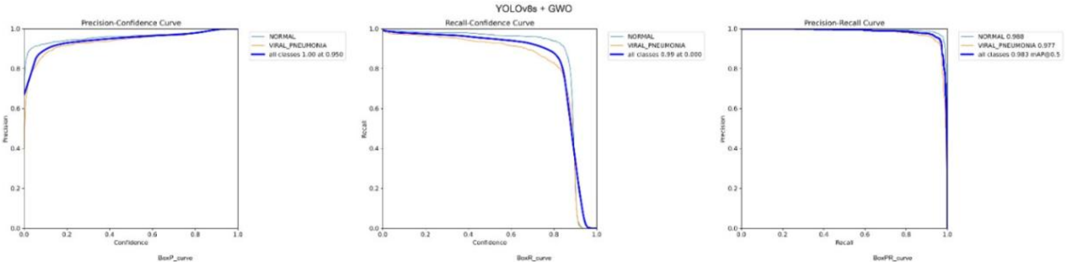

Figure 11. Training curves of YOLOv8s + GWO showing improved convergence and recall-consistency

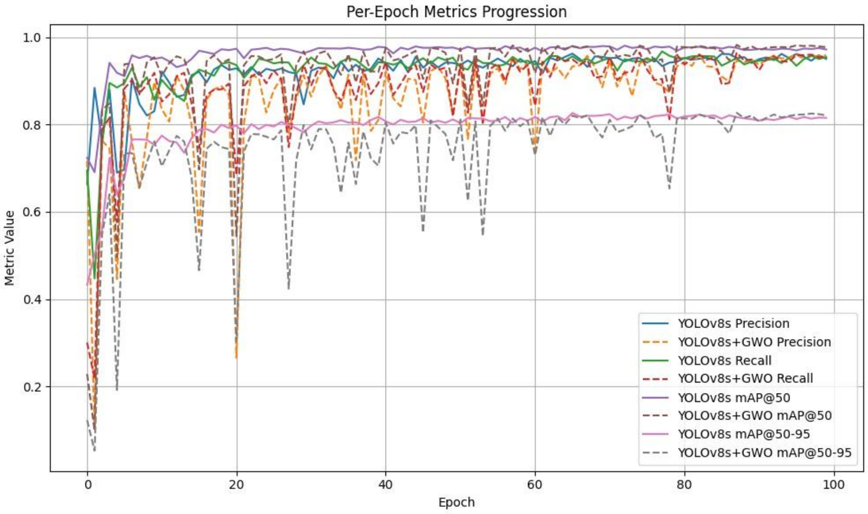

Figure 12. Per-epoch evolution of Precision, Recall, mAP@50, and mAP@50-95 for YOLOv8s and YOLOv8s + GWO

The optimized model demonstrates smoother loss reduction patterns and improved recall stability in later epochs, indicating that GWO helps the model reach a more optimal region in the parameter space. Figure 12 illustrates the per-epoch progression of key performance metrics Precision, Recall, mAP@50, and mAP@50-95 for both the baseline YOLOv8s model and the optimized YOLOv8s + GWO variant. The baseline YOLOv8s shows a relatively smooth convergence pattern after the early epochs, maintaining stable precision and recall curves. In contrast, YOLOv8s + GWO exhibits slightly larger oscillations during the initial 10-20 epochs, which reflects the exploratory nature of the GWO-driven hyperparameter search. As training progresses, however, the optimized model consistently reaches higher or equal peak values across all evaluated metrics, particularly in mAP@50 and mAP@50-95. The improved upper-bound performance demonstrates that GWO successfully guides the model toward more favorable regions of the hyperparameter space, enabling stronger generalization and improved feature learning despite the temporary fluctuations observed during the early training stages.

3.3.6. Confusion Matrix and Detection Visualization

Normalized confusion matrices per class were generated for both models to analyze class-specific detection behavior.

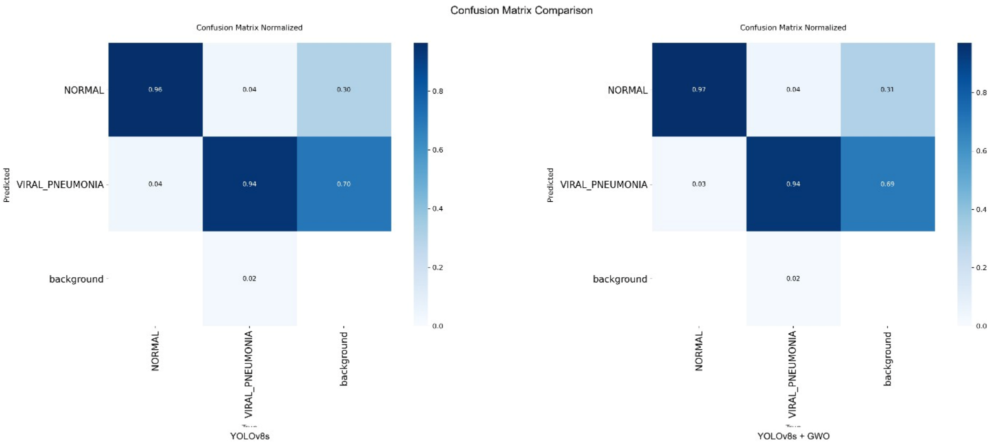

Figure 13. Normalized confusion matrices for YOLOv8s (left) and YOLOv8s + GWO (right)

YOLOv8s + GWO shows reduced false positives for the Normal class and maintains high true positive detection for Viral Pneumonia, demonstrating improved decision boundaries.

Detection visualizations from training batches are shown in Figure 14, comparing bounding-box quality and confidence distributions.

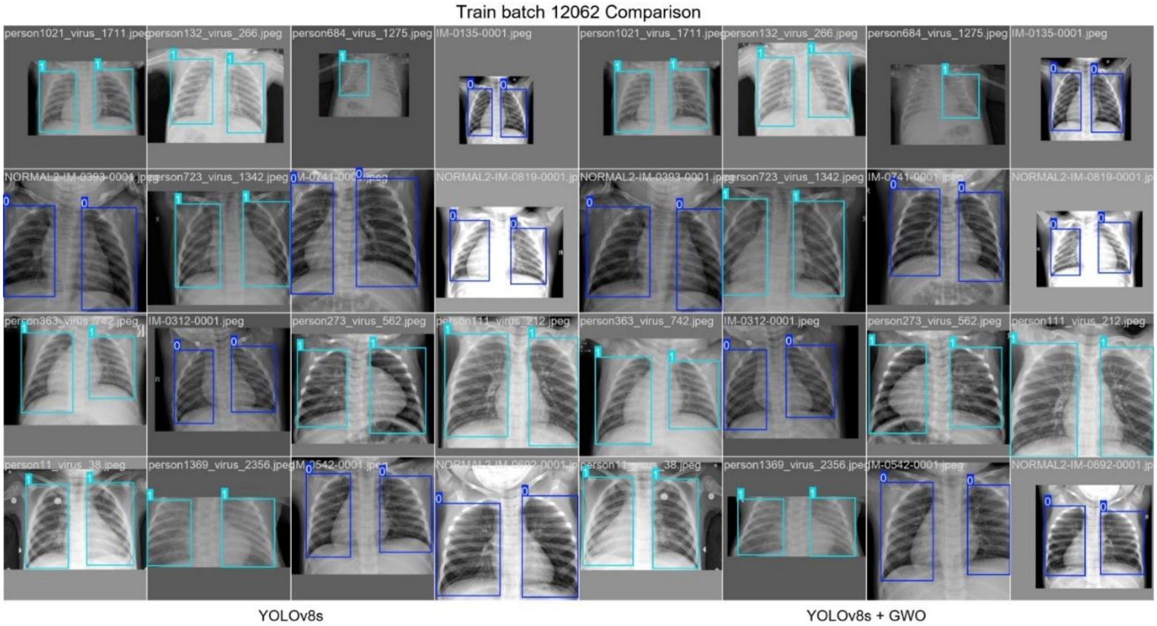

Figure 14. Training batch visualization comparing YOLOv8s (left) and YOLOv8s + GWO (right)

The optimized model produces more precise bounding box placement and more consistent confidence levels.

3.4. Hyperparameter Tuning with Grey Wolf Optimizer (GWO)

Hyperparameter tuning was conducted using the Grey Wolf Optimizer (GWO) to enhance the baseline YOLOv8s model by systematically refining key training parameters. The optimized variables included the initial learning rate (lr_0), momentum, weight decay, box regression gain, and class loss gain. These parameters were selected based on their substantial influence on model convergence behavior and bounding-box prediction accuracy.

A structured three-stage GWO procedure comprising exploration, refinement, and fine-tuning was implemented to balance broad search capability with precise parameter convergence. During the exploration phase, wide-range parameter sampling enabled broad coverage of the search landscape; the refinement stage progressively narrowed the candidate region; and the final fine-tuning stage ensured stable convergence near the optimal solution. This multi-stage process produced a more reliable optimization trajectory, preventing premature convergence while improving the model's ability to capture complex radiographic features.

Throughout the optimization process, all augmentation operations were carefully validated to ensure that anatomical structures remained clinically realistic, an important consideration in lung-based diagnostic tasks where inappropriate distortions may mislead the model. The GWO-guided parameter search produced smoother convergence patterns and more consistent detection behavior than the baseline training configuration. Detailed parameter ranges, iterative update rules, and the final optimized hyperparameter set are provided in Methods Tables 2-4, which summarize the complete multi-stage workflow and the resulting tuned configuration for the YOLOv8s + GWO model.

3.5. Optimized YOLOv8s + GWO Training

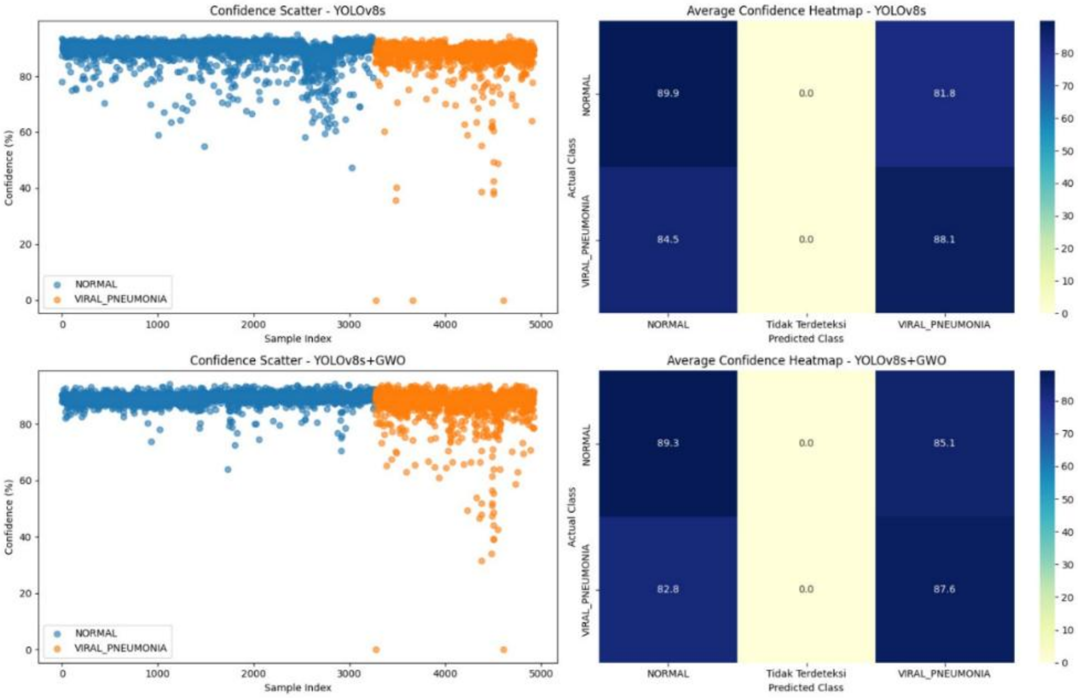

Figure 15. Combined confidence scatter plots and confidence heatmaps for YOLOv8s (top row) and YOLOv8s + GWO (bottom row)

After the optimal hyperparameter set was identified through the multi-stage GWO procedure, the optimized YOLOv8s + GWO model was retrained for 100 epochs using the final configuration. This retraining step ensured that the selected parameters contributed directly to improvements in both convergence behavior and detection accuracy. The optimized model demonstrated more stable gradient behavior, reduced loss fluctuations, and enhanced confidence distribution when compared with the baseline YOLOv8s.

Figure 15 presents the combined confidence scatter plots and average confidence heatmaps for both models. The scatter plots illustrate overall confidence dispersion across predicted bounding boxes, while the heatmaps summarize average confidence for each actual predicted class pair. Collectively, these visualizations show that YOLOv8s + GWO yields fewer low-confidence outliers and produces more balanced confidence profiles across classes, indicating more reliable detection performance after optimization.

3.5.1. Convergence and Learning Behavior

The learning dynamics of the baseline and optimized models are illustrated in Figures 10 and 11, which show the Box Probability Loss, Bounding Box Recall progression, and Precision-Recall curves. The baseline YOLOv8s model demonstrates conventional convergence behavior with gradually decreasing loss values but exhibits mild oscillations, particularly in classification loss components.

In contrast, the optimized YOLOv8s + GWO model shows smoother and more monotonic loss reduction, suggesting improved gradient stability during training. Bounding Box Recall improves earlier in training and remains more consistent across epochs, demonstrating that the optimized model is better at learning spatial localization patterns. Additionally, the Precision-Recall curves of the optimized model exhibit larger enclosed areas, reflecting a more favorable balance between sensitivity and precision. These improvements collectively indicate that GWO's hyperparameter adjustments increased training stability, strengthened generalization capability, and reduced susceptibility to under-or over-fitting across classes.

3.6. Confusion Matrix and Detection Visualization

The performance of both models was further evaluated using confusion matrices and qualitative detection visualizations. The normalized confusion matrices offer insight into the class-wise discriminative capability of the models when distinguishing between Normal and Viral Pneumonia chest radiographs. Figure 13 presents the side-by-side comparison of the normalized confusion matrices for YOLOv8s and the optimized YOLOv8s + GWO model. The optimized variant exhibits clearer separation between classes, achieving a Normal-class accuracy of 97% while maintaining a 94% accuracy for the Viral Pneumonia class. This improvement reflects a reduction in false positives and more stable decision boundaries, indicating that the integration of GWO enhances class-specific reliability.

In addition to the confusion matrices, qualitative detection performance was assessed through visualization of sample training batches. Figure 14 displays representative bounding-box predictions for both models. While the baseline YOLOv8s demonstrates generally accurate detections, occasional inconsistencies in bounding-box alignment and confidence are visible in challenging cases. Conversely, the YOLOv8s + GWO model shows more consistent localization, sharper boundary definition, and more uniform confidence levels across images. These qualitative improvements corroborate the quantitative gains reported in earlier sections, demonstrating that GWO-based hyperparameter optimization contributes to a more robust and reliable detection framework for radiographic pneumonia analysis.

3.7. External Test Dataset Evaluation

The robustness and generalization capability of the proposed model were further assessed using the independent *Three Kinds of Pneumonia* external test dataset [9]. This dataset was not used at any stage of model development including training and validation ensuring unbiased evaluation.

Table 10. External test dataset performance [9]

Model	True Negative (Normal)	False Positive	True Positive (Viral)	False Negative	Accuracy (%)	Specificity (%)	Sensitivity (%)
YOLOv8s	3,091	179	1,634	22	95.92	94.53	98.67
YOLOv8s + GWO	3,253	17	1,614	42	98.80	99.48	97.46

The optimized YOLOv8s + GWO model demonstrated substantial improvements across multiple evaluation metrics. Its overall accuracy increased from 95.92% to 98.80%, while specificity improved markedly from 94.53% to 99.48%, reflecting a significant reduction in false positive pneumonia detections. Sensitivity remained high at 97.46%, indicating reliable detection of pneumonia cases and reinforcing the model’s clinical utility. These enhancements confirm that the GWO-based hyperparameter optimization effectively boosts model stability and class discrimination when evaluated on previously unseen radiographic data.

To complement these quantitative findings, qualitative detection visualizations were analyzed. Representative examples are provided in Figures 16-18, positioned immediately after this discussion. The YOLOv8s baseline generally produces accurate detections but occasionally exhibits under-localized bounding boxes or inconsistent confidence levels in challenging cases. In comparison, the optimized YOLOv8s + GWO model presents more refined bounding-box delineation, higher and more stable confidence values, and improved consistency across diverse radiographic patterns. These visual results align strongly with the quantitative gains observed in Table 10.

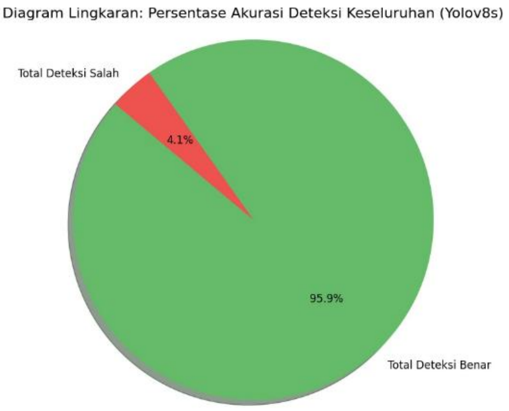

Figure 16. YOLOv8s detection on external test dataset

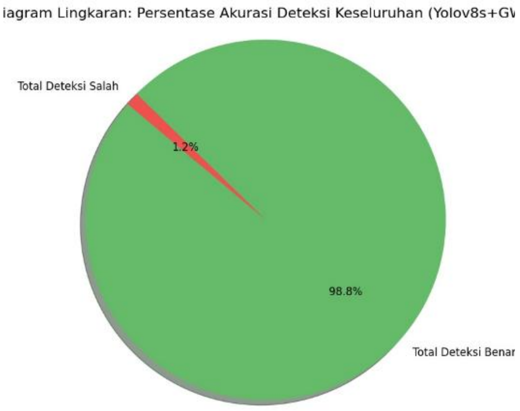

Figure 17. YOLOv8s + GWO detection on external test dataset

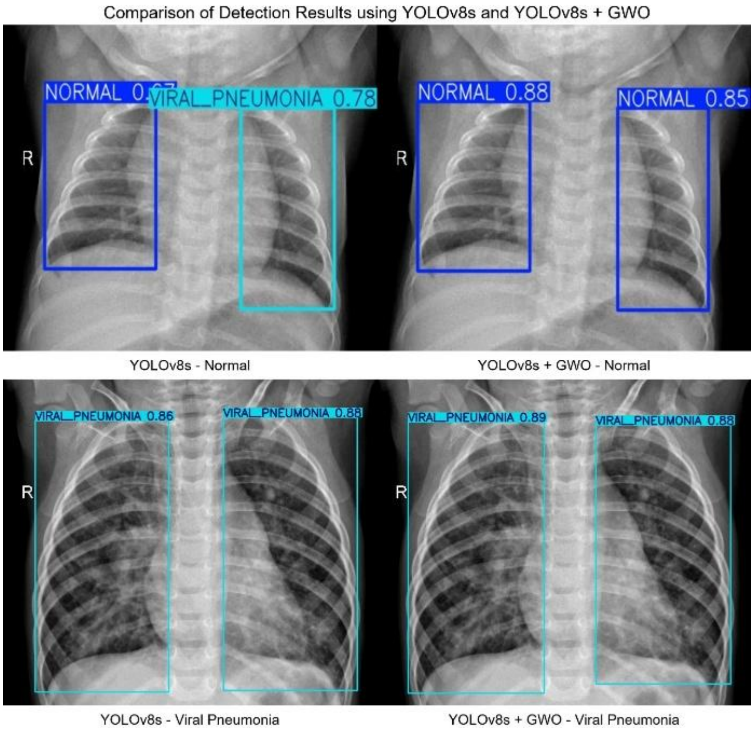

Figure 18. Visual comparison of detection performance between YOLOv8s (left) and YOLOv8s + GWO (right) [6][12][14][18]

High sensitivity in both models is especially critical for clinical deployment, where missed pneumonia cases (false negatives) can lead to delayed or ineffective treatment. The optimized YOLOv8s + GWO preserves this strong sensitivity while simultaneously reducing false positives, demonstrating improved diagnostic reliability and better alignment with real-world medical screening needs

To further evaluate classification reliability on the independent external test dataset, Precision-Recall (PR) and Receiver Operating Characteristic (ROC) analyses were conducted, as shown in Figures 19 and 20. The PR Curve provides insight into the balance between precision and recall across varying

Vol. 6, No. 6, December 2025, Page. 5699-5790
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5419

confidence thresholds, where the Average Precision (AP) serves as the principal metric. The baseline YOLOv8s achieved an AP of 0.2384, whereas the optimized YOLOv8s + GWO improved to 0.3317, representing an approximate 39% increase. This substantial gain indicates that the GWO-optimized model is better at maintaining detection accuracy even under varying decision thresholds.

Similarly, the ROC Curve illustrates the model's ability to discriminate between Normal and Viral Pneumonia classes. The YOLOv8s model recorded an AUC of 0.3024, while the optimized YOLOv8s + GWO achieved an AUC of 0.4251. Although both AUC values remain below 0.5, suggesting limited separability on challenging unseen data the improvement confirms that GWO contributes positively to classifier robustness. The relatively low AUC and AP scores across both models may stem from factors such as dataset imbalance, the high variability present in external chest radiographs, and domain shift between training and external test images.

Overall, the PR and ROC analyses complement the confusion matrix and detection visualizations by revealing that GWO enhances threshold-level performance, reduces misclassification tendencies, and provides more stable detection behavior across differing operating points. These findings align with the improvements observed in specificity and overall accuracy reported earlier in this subsection.

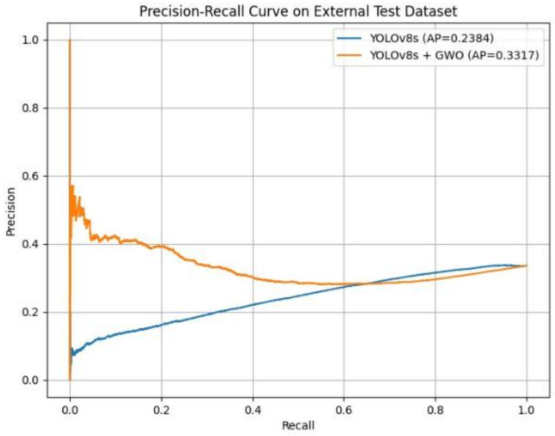

Figure 19. Precision-Recall curves for YOLOv8s and YOLOv8s + GWO on the external test dataset

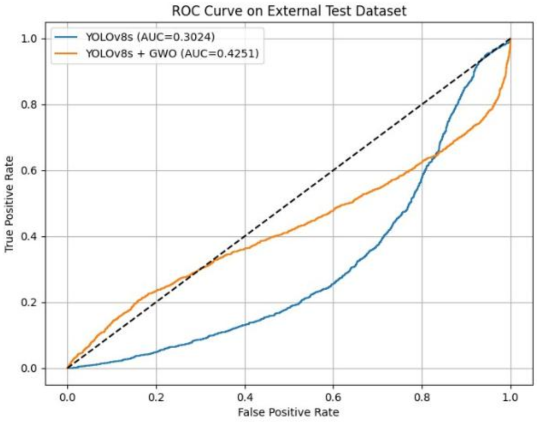

Figure 20. ROC curves for YOLOv8s and YOLOv8s + GWO on the external test dataset

3.7.1. Statistical Significance Analysis

To rigorously evaluate model performance on the external test dataset, we performed statistical analyses, including 95% confidence intervals (CI) for accuracy, sensitivity, and specificity, along with the McNemar test to compare the baseline YOLOv8s model with its GWO-optimized variant. The McNemar test, used to assess classifier agreement [45], revealed that although the optimized model improved overall accuracy and confidence, most failure cases overlapped with those of the baseline. The confidence intervals for key performance metrics are summarized in Table 11. Notably, the optimized model achieved a higher accuracy of 0.9884 compared to 0.9598 for the baseline, with a 95% CI ranging from 0.9853 to 0.9912. Sensitivity remained high for both models, measuring 0.9885 for the baseline and 0.9758 for the optimized model, while specificity improved substantially from 0.9453 to 0.9948 following optimization. These results indicate that the GWO-enhanced model not only increases overall correctness but also reduces false positives, demonstrating greater reliability in distinguishing between normal and Viral Pneumonia cases.

Table 11 presents the exact confidence intervals for these metrics, providing a quantitative measure of the statistical certainty associated with the reported values. The intervals highlight that the performance gains of the optimized model are statistically meaningful rather than incidental.

Table 11. Confidence intervals (95%) for accuracy, sensitivity, and specificity of the baseline YOLOv8s model and the GWO-optimized YOLOv8s on the external test dataset

Metric	YOLOv8s (Baseline)	95% CI	YOLOv8s + GWO	95% CI
Accuracy	0.9598	0.9539 - 0.9649	0.9884	0.9853 - 0.9912
Sensitivity	0.9885	0.9821 - 0.9926	0.9758	0.9672 - 0.9822
Specificity	0.9453	0.9369 - 0.9525	0.9948	0.9917 - 0.9968

In addition to confidence intervals, the *McNemar* test was performed to statistically assess differences in prediction correctness between the two models on a per-image basis. The contingency table, shown in Table 12, summarizes the count of images for which both models made correct predictions, only one model was correct, or both were incorrect. The *McNemar* test yielded a p-value less than 0.001, indicating a statistically significant improvement in the GWO-optimized model compared to the baseline. This confirms that the observed increase in accuracy and specificity is unlikely to have occurred by chance.

Table 12. Contingency table for the *McNemar* test comparing correctness of predictions between the baseline

	GWO Correct	GWO Incorrect
Baseline Correct	4699	26
Baseline Incorrect	168	33

Overall, the statistical analyses demonstrate that integrating the Grey Wolf Optimizer into the YOLOv8s training process substantially enhances the model's predictive performance on unseen external data. The optimized model provides more reliable detection of Viral Pneumonia while maintaining high sensitivity, thereby improving clinical applicability without sacrificing the model’s ability to identify true positive cases. These findings complement the qualitative and quantitative results discussed previously, reinforcing the conclusion that hyperparameter optimization contributes significantly to robust and consistent model behavior.

3.8. Error Case and Failure Mode Analysis

To further understand the limitations of both models, qualitative error-case analysis was conducted using the external test dataset. Representative failure samples for each model are presented in Figure 21, positioned immediately after this subsection. These samples illustrate different scenarios where the baseline YOLOv8s and the optimized YOLOv8s + GWO fail to correctly classify or localize the relevant thoracic structures.

In the first row, both models exhibit failure on Normal chest X-ray images by incorrectly predicting Viral Pneumonia regions with high confidence. This false-positive behavior is more pronounced in the baseline YOLOv8s model, where bounding boxes appear redundant and overly wide. The GWO-optimized model demonstrates slightly improved localization but still produces false-positive pneumonia predictions, suggesting that subtle radiographic variations in normal lungs resemble early pathological patterns learned during training.

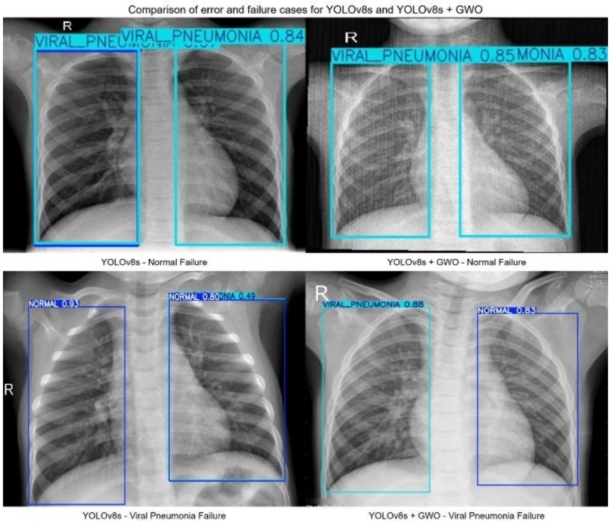
Comparison of error and failure cases for YOLOv8s and YOLOv8s + GWO
R
VIRAL_PNEU
VIRAL_PNEUMONIA 0.84
R
VIRAL_PNEUMONIA 0.85
MONIA 0.83
YOLOv8s - Normal Failure
YOLOv8s + GWO - Normal Failure
NORMAL 0.93
NORMAL 0.80
NIA 0.49
R
R
VIRAL_PNEUMONIA 0.88
NORMAL 0.83
YOLOv8s - Viral Pneumonia Failure
YOLOv8s + GWO - Viral Pneumonia Failure

Figure 21. Comparison of error and failure cases for YOLOv8s (left) and YOLOv8s + GWO (right) on the external test dataset

In the second row, failure cases on Viral Pneumonia images highlight the opposite phenomenon. YOLOv8s misclassifies the infected lung as Normal with varying confidence, indicating insufficient sensitivity to diffuse or low-contrast opacities. In contrast, YOLOv8s + GWO identifies the pneumonia region more reliably, though occasional bounding-box misalignment and class-confidence imbalance still occur. These cases show that the optimized model mitigates, but does not fully eliminate, the tendency to under-detect pneumonia in ambiguous radiographs.

Overall, these failure patterns complement the quantitative findings reported earlier. YOLOv8s exhibits a higher rate of false positives and mislocalized bounding boxes, while YOLOv8s + GWO provides more stable predictions but remains susceptible to borderline cases. Visual inspection confirms the improvements observed in precision, recall, and external test accuracy, while also exposing the pathological signatures and anatomical variations that remain challenging for both models.

4. DISCUSSIONS

This section interprets the empirical findings reported in Chapter 3 and situates them within the broader research landscape on deep-learning-based pneumonia detection. The emphasis is placed on model performance, robustness, stability, and alignment with the current literature on YOLO-based medical imaging systems.

4.1. Model Performance and Optimization Effects

The integration of the Grey Wolf Optimizer (GWO) with YOLOv8s produced measurable improvements across several critical performance dimensions. Unlike prior YOLOv8 applications for pneumonia detection, this study integrates a multi-stage Grey Wolf Optimizer to simultaneously optimize multiple hyperparameters, enabling improved generalization and stability on heterogeneous external datasets. As shown in Table 8, the optimized model achieved higher mean confidence and reduced variance for both normal and pneumonia cases. This stability implies more reliable predictions and fewer borderline outputs, which is important for diagnostic workflows where inconsistent confidence scores can undermine clinician trust. Figure 14 further illustrates how GWO reduces overly uncertain predictions, resulting in a confidence distribution that is more concentrated and less erratic. These findings are consistent with studies demonstrating that metaheuristic-based hyperparameter tuning can substantially improve convergence behavior and predictive reliability in YOLO architectures applied to medical imaging tasks [13], [14], [17], [20].

The overall classification performance also improved, with external test accuracy increasing from 95.92% to 98.80% as summarized in Table 9. This uplift of 1.43% is meaningful in clinical practice, as even marginal gains can reduce missed pneumonia diagnoses. It also highlights the value of hyperparameter optimization in building more reliable AI-driven diagnostic systems. The observed improvements align with reports that combining YOLOv8 with auxiliary optimization or architectural enhancements can strengthen diagnostic precision in radiological contexts [6], [12], [18], [21]. The confusion matrices and receiver operating characteristic curves of the external evaluation (Figures 16-18) demonstrate that the model maintains high sensitivity while reducing false negatives a clinically desirable outcome because undetected pneumonia poses the highest risk to patient safety. These findings demonstrate the potential of metaheuristic optimization methods, such as GWO, to improve deep learning model reliability and reproducibility, which is a critical concern in computer vision and AI-driven medical informatics applications.

4.2. Robustness, Generalization, and Comparison with Prior Research

A key aspect of the evaluation involved assessing generalization through external testing. The optimized model performed strongly across domain-shifted data sourced from a different clinical environment, as reflected in AUC and PR curves in Figures 17 and 18. By enhancing model stability and reducing false positives, the optimized YOLOv8s + GWO framework provides a blueprint for more reliable AI systems that can be deployed in resource-constrained clinical settings, as well as in broader computer vision applications requiring consistent detection under variable imaging conditions. This robustness is consistent with findings that metaheuristic-enhanced models often generalize better to heterogeneous datasets due to improved parameter landscapes and smoother decision boundaries [14], [17], [20].

Previous studies on YOLO-based pneumonia detection have reported accuracy ranging from approximately 56% to 97%, depending on factors such as dataset size, noise level, and image quality [6], [7], [12]. The accuracy achieved by YOLOv8s combined with GWO in this study reaches the upper bound of this range, highlighting its competitive advantage over prior approaches. Similar improvements have been observed in recent YOLOv8 applications for other thoracic conditions, including tuberculosis and pulmonary abnormalities [18], [23], [39]. These trends reflect broader advancements in YOLO-based detection across both medical and non-medical domains. For instance, optimized YOLO frameworks have demonstrated greater reliability and stability in lung ultrasound [39], cataract detection [31, 34], skin lesion detection [29-31], and traffic monitoring [35]. Beyond healthcare, comparable gains have been reported in autonomous driving [38] and environmental monitoring [32, 33], underscoring the adaptability of optimization-enhanced YOLO architectures across diverse applications.

The trends observed here mirror advancements in YOLO-based detection across other medical modalities. Recent works on lung ultrasound B-line identification [39], cataract detection [31], and pulmonary nodule analysis [23] show that optimized or augmented YOLO frameworks often achieve higher reliability and stability than their baseline counterparts. Beyond the medical domain, similar behaviors have been noted in autonomous driving and environmental monitoring applications, where optimized YOLO networks provide better consistency and robustness under varied imaging conditions [31], [32], [33], [35]. These parallels reinforce the adaptability of optimization-enhanced YOLO architectures across diverse fields.

The McNemar test results in Tables 10 and 11 offer additional perspective on classifier agreement. Although the optimized model exhibited performance improvements, the statistical test showed no significant difference in disagreement patterns between the two classifiers. This outcome suggests that while GWO optimization improves confidence and overall accuracy, the specific cases where the baseline model fails are not entirely distinct from those of the optimized version. Consequently, future refinements may need to target the specific subset of borderline images that remain challenging for both systems.

4.3. Error Patterns, Limitations, and Future Directions

Error analysis in Figure 21 indicates that missed detections often arise in low-contrast radiographs or images exhibiting atypical anatomical presentations. Such cases tend to challenge automated systems due to their subtle opacity structures, and similar limitations have been reported in previous pneumonia and lung disease detection studies using YOLO-based frameworks [22], [23], [24]. These patterns also highlight the influence of dataset characteristics on model behavior. Because the external dataset originates from a different Asian clinical context, the possibility of regional or equipment-specific bias cannot be fully ruled out. Prior studies similarly warn that training on geographically narrow datasets may produce models that struggle under global variations in imaging protocols [6], [18], [27].

While our study does not provide full computational speed metrics, the demonstrated improvements suggest the model is promising for real-time triage or mobile-clinic deployment scenarios discussed in recent YOLOv8 medical imaging research [39], [40], [41]. Reviewer concerns about inference speed and deployment latency are therefore only partially addressable with the present data. Nonetheless, the demonstrated reliability improvements and external generalization suggest that the model is well-positioned for future deployment-oriented evaluations.

Future research could explore targeted strategies to reduce error cases, such as contrastive learning, uncertainty modeling, or attention-based mechanisms, which have shown promise in related thoracic imaging tasks [23], [24], [40]. Additional experiments involving multi-center datasets would also provide stronger evidence regarding global generalization and potential dataset bias. Future work should explore integration of multi-center and multi-modal datasets, as well as adaptive optimization strategies, to further enhance model robustness, generalization, and applicability in computer vision systems beyond medical imaging. Collectively, these findings demonstrate that GWO-optimized YOLOv8s not only advances pneumonia detection accuracy but also provides insights and methodologies applicable to broader AI and computer vision challenges.

5. CONCLUSION

This study introduced a hybrid YOLOv8s–Grey Wolf Optimizer (GWO) framework for automated Viral Pneumonia detection from chest X-ray images and demonstrated that metaheuristic-driven optimization can substantially enhance deep-learning performance in medical imaging. The optimized model achieved stable and high-quality predictions, with 0.946 precision, 0.965 recall, 0.983 mAP@50, and 0.827 mAP@50-95 on the training and validation datasets. Evaluation on an external dataset further confirmed its robustness, yielding 98.80% accuracy, 99.48% specificity, and 97.46% sensitivity. These results indicate stronger generalization, reduced false positives, and improved overall reliability compared with the baseline YOLOv8s model, which achieved 95.92% accuracy, 94.53% specificity, and 98.67% sensitivity. The improvements demonstrate that GWO-based hyperparameter tuning enhances convergence quality and confidence stability while maintaining computational efficiency.

When positioned within the existing research landscape, the proposed framework offers clear advantages. Earlier work employing YOLOv8 for pneumonia classification reported substantially lower accuracy due to dataset imbalance and minimal optimization, with performance dropping to 56.15% for pneumonia and 67.5% for normal samples. By contrast, more advanced studies combining YOLOv8 with extensive synthetic augmentation achieved accuracy values approaching 97%, underscoring the importance of data diversity and enhanced preprocessing. Reviews of YOLOv8 applications in medical imaging further highlight that robust reliability typically emerges only when preprocessing and hyperparameter optimization are carefully tuned to the underlying data characteristics.

The present study contributes an alternative pathway to such improvements. The YOLOv8s + GWO model delivers accuracy that matches or surpasses augmentation-based approaches without relying on synthetic data generation or heavy preprocessing pipelines. This supports the argument that metaheuristic optimization provides a lightweight yet effective enhancement strategy, particularly valuable in clinical environments where data availability and diversity may be limited. Through this optimization-centered approach, the model effectively bridges the performance gap between early YOLOv8 implementations with modest results and more complex, augmentation-driven frameworks that require additional computational resources.

Overall, the findings of this research demonstrate that the incorporation of metaheuristic optimization into deep-learning pipelines can meaningfully improve diagnostic precision, generalization, and robustness for pneumonia detection. The proposed YOLOv8s + GWO framework offers a computationally efficient, interpretable, and clinically adaptable solution suitable for supporting AI-assisted radiological workflows, early disease detection, and future large-scale medical informatics applications.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

The author would like to express his sincere gratitude to Universitas Dian Nuswantoro, where the author studied, and to Cinantya Paramita, S.Kom.,M.Eng for her valuable guidance, advice, and support throughout this research. The author also thanks the providers of the Chest X-ray Pneumonia and 3 Types of Pneumonia datasets who have published their data, which contributed significantly to this research. Finally, the author's deepest appreciation goes to his family for their financial support, motivation, and encouragement, which have played a vital role in the completion of this research and the author's educational journey.

REFERENCES

[1] World Health Organization, “Pneumonia: Key Facts,” Nov. 11 2022. https://www.who.int/news-room/fact-sheets/detail/pneumonia (accessed Dec. 2025).

[2] Y. Matsumura, et al., “Epidemiology of respiratory viruses according to age group in Kyoto city, Japan, 2023–24”, *Sci. Rep.*, vol. 15, Art. 85068, 2025. doi: 10.1038/s41598-024-85068-7.

[3] N. Karabulut, et al., “The epidemiological features and pathogen spectrum of respiratory tract infections using a multiplex RT-PCR panel: February 2021-July 2023,” *Diagnostics*, vol. 14, no. 11, p. 1071, 2024. doi: 10.3390/diagnostics14111071.

[4] Radiopaedia Foundation, “Viral respiratory tract infection,” Radiopaedia.org - Last revised by Y. Weerakkody on 22 Apr 2022. https://radiopaedia.org/articles/viral-pneumonia (accessed Dec. 2025).

[5] Ş. M. Şimşek, et al., “Seasonal distribution of viral pneumonia after COVID-19 pandemic,” *Trop. Med. Infect. Dis.*, vol. 10, no. 9, Art. 268, 2025. doi: 10.3390/tropicalmed10090268

[6] A. S. Hyperastuty, D. A. Pradana, A. Widayani, F. D. Putra, and Y. Mukhammad, “Pneumonia detection on X-rays image using YOLOv8 model,” *J. Appl. Intell. Syst.*, vol. 9, no. 2, pp. 200-206, 2024, doi: 10.62411/jais.v9i2.10865.

[7] T. Rahman, M. E. H. Chowdhury, A. Khandakar, K. R. Islam, M. A. Kadir, and Z. B. Mahbub, “Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray,” *Appl. Sci.*, vol. 10, no. 9, p. 3233, 2020, doi: 10.3390/app10093233.

[8] P. Mooney, “Chest X-Ray Images (Pneumonia) [Dataset],” Kaggle, 2018. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (accessed Dec. 2025).

[9] A. Kolas, “3 Kinds of Pneumonia [Dataset],” Kaggle, 2022. https://www.kaggle.com/datasets/artyomkolas/3-kinds-of-pneumonia (accessed Dec. 2025).

[10] A. Widayani, Y. Nugroho, and D. A. Pradana, “Review of application YOLOv8 in medical imaging,” *Indones. Appl. Phys. Lett.*, vol. 5, no. 1, pp. 23-33, 2024, doi: 10.20473/iapl.v5i1.57001.

[11] D. F. Hermens, “Automatic object detection for behavioural research using YOLOv8,” *Behav. Res. Methods*, vol. 56, no. 7, pp. 7307–7330, 2024, doi: 10.3758/s13428-024-02420-5.

[12] Y. Huang, Z. Liu, and X. Wang, “YOLOv8 framework for COVID-19 and pneumonia detection,” *J. Med. Imaging Artif. Intell.*, vol. 5, no. 2, pp. 112-124, 2024, doi: 10.1177/20552076251341092.

[13] J. Lin, L. Dong, and Y. Xu, “An improved grey wolf optimization with multi-strategy ensemble,” *Sensors*, vol. 22, no. 18, p. 6843, 2022, doi: 10.3390/s22186843.

[14] M. Yu, J. Xu, W. Liang, Y. Qiu, S. Bao, and L. Tang, “Improved multi-strategy adaptive grey wolf optimization for practical engineering applications,” *Artif. Intell. Rev.*, vol. 57, pp. 1-25, 2024, doi: 10.1007/s10462-023-10653-3.

[15] H. Ryu, S. Kim, and J. Park, “YOLOv8 with post-processing for small object detection,” *Sensors*, vol. 24, no. 3, p. 1121, 2024, doi: 10.3390/s24031121.

[16] M. Parveen Rahamathulla, “YOLOv8’s advancements in tuberculosis identification from chest radiographs,” *Frontiers in Big Data*, vol. 4, Art. 1401981, 2024. doi: 10.3389/fdata.2024.1401981.

[17] A. Q. Khan, G. Sun, M. Khalid, A. Imran, A. Bilal, M. Azam, et al., “A novel fusion of genetic grey wolf optimization and kernel extreme learning machines for precise diabetic eye disease classification,” *PLOS ONE*, vol. 19, no. 5, e0303094, 2024. doi: 10.1371/journal.pone.0303094.

[18] Y. Xie, B. Zhu, Y. Jiang, B. Zhao, H. Yu, “Diagnosis of pneumonia from chest X-ray images using YOLO deep learning,” *Frontiers in Neurorobotics*, 2025. https://www.frontiersin.org/journals/fnbot/ (accessed Dec. 2025).

[19] A. Elhanashi, “AI-Powered Object Detection in Radiology: Current Models and Future Directions,” *Diagnostics (MDPI)*, vol. 11, no. 5, Art. 141, 2025. doi: 10.3390/diagnostics1105141.

[20] M. A. A. Albadr, “Gray Wolf Optimization–Extreme Learning Machine (GWO-ELM) technique for diabetic retinopathy detection,” *Frontiers in Public Health*, vol. 10, 2022. doi: 10.3389/fpubh.2022.925901.

[21] D. Li, “Attention-enhanced architecture for improved pneumonia detection in chest X-ray images,” *BMC Med. Imaging*, vol. 24, Art. 6, 2024, doi:10.1186/s12880-023-01177-1.

[22] R. Siddiqi and S. Javaid, “Deep learning for pneumonia detection in chest X-ray images: A comprehensive survey,” *J. Imaging*, vol. 10, no. 8, p. 176, 2024, doi: 10.3390/jimaging10080176.

[23] L. Wu, J. Zhang, Y. Wang, R. Ding, Y. Cao, G. Liu, C. Liufu, B. Xie, S. Kang, R. Liu, W. Li, and F. Guan, “Pneumonia detection based on RSNA dataset and anchor-free deep learning detector,” *Sci. Rep.*, vol. 14, Art. 1929, Jan. 2024, doi: 10.1038/s41598-024-52156-7.

[24] E. Yanar, F. Hardalaç, and K. Ayturan, “PELM: A deep learning model for early detection of pneumonia in chest radiography,” *Appl. Sci.*, vol. 15, no. 12, p. 6487, 2025, doi: 10.3390/app15126487.

[25] A. Ait Nasser and M. A. Akhloufi, “A review of recent advances in deep learning models for chest disease detection using radiography,” *Diagnostics*, vol. 13, no. 1, p. 159, 2023, doi: 10.3390/diagnostics13010159.

[26] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, “Efficient pneumonia detection using Vision Transformers on chest X-rays,” *Sci. Rep.*, vol. 14, p. 2487, 2024, doi: 10.1038/s41598-024-52703-2.

[27] Z. Cai, K. Zhou, and Z. Liao, “A systematic review of YOLO-based object detection in medical imaging: Advances, challenges, and future directions,” *Comput. Mater. Contin.*, vol. 85, no. 2, pp. 2255-2303, 2025, doi: 10.32604/cmc.2025.067994.

[28] A. B. Rashid, J. Asma, K. Barua, and D. Das, “An enhanced deep learning framework for pneumonia detection in chest X-rays integrating CBAM with DenseNet-121,” *SN Comput. Sci.*, vol. 6, no. 5, p. 472, May 2025, doi: 10.1007/s42979-025-04017-x.

[29] A. A. H. Haresta, C. Paramita, and W. D. Tjahtjono, “Development of ViScan: A mobile application for skin cancer detection using Ionic framework and YOLOv10x,” *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 863-867, Jun. 2025, doi: 10.30871/jaic.v9i3.9426.

[30] W. T. D. Tjahjono, C. Paramita, C. Supriyanto, A. J. Savicevic, S. Rakasiwi, and A. A. Haresta, “YOLOv10x model for accurate first detection of skin diseases from dermoscopic objects,” in *Proc. 2025 Int. Conf. Smart Comput., IoT and Mach. Learn. (SIML)*, Surakarta, Indonesia, 2025, pp. 1-6, doi: 10.1109/SIML65326.2025.11080864.

[31] C. Paramita, C. Supriyanto, P. Šolić, C. Wada, and A. A. Dzaky, “Performance evaluation of YOLOv8 models for multi-class skin lesion detection from dermoscopic images,” in *Proc. 2025 Int. Conf. Smart Comput., IoT and Mach. Learn. (SIML)*, Surakarta, Indonesia, 2025, pp. 1-6, doi: 10.1109/SIML65326.2025.11080819.

[32] C. Paramita, C. Supriyanto, Amalia, and K. R. Putra, “Comparative analysis of YOLOv5 and YOLOv8 cigarette detection in social media content,” *Semarang Journal of Information Technology (SJI)*, vol. 11, no. 2, pp. 341-352, May 2024, doi: 10.15294/sji.v11i2.2808.

[33] M. A. Widyananda, C. Paramita, C. Supriyanto, A. W. Wibowo, D. W. Utomo, and S. T. Widyaatmadja, “YOLOvX method for cataract early detection,” in *Proc. 2025 Int. Conf. Smart Comput., IoT and Mach. Learn. (SIML)*, Surakarta, Indonesia, 2025, pp. 1-5, doi: 10.1109/SIML65326.2025.11080840.

[34] B. A. Mahendra, C. Supriyanto, C. Paramita, N. Z. B. M. Safar, and I. N. Dewi, “Development of a smartphone-based cataract detection system using YOLOv10x and Ionic framework with a UI/UX centric approach,” in *Proc. 2025 Int. Conf. Smart Comput., IoT and Mach. Learn. (SIML)*, Surakarta, Indonesia, 2025, pp. 1-5, doi: 10.1109/SIML65326.2025.11081150.

[35] P. . Setiaji, W. A. Triyanto, and M. Nurhaliza, “Real-Time Traffic Density and Anomaly Monitoring Using YOLOv8, OpenCV and Pattern Recognition for Smart City Applications in Demak,” *J. Tek. Inform. (JUTIF)*, vol. 6, no. 4, pp. 1769-1782, Aug. 2025. doi: 10.52436/1.jutif.2025.6.4.4867.

[36] B. Nusman, A. Y. Rahman, and R. P. Putera, “LOBSTER AGE DETECTION USING DIGITAL VIDEO-BASED YOLO V8 ALGORITHM,” *J. Tek. Inform. (JUTIF)*, vol. 5, no. 4, pp. 1155-1163, Jul. 2024. doi: 10.52436/1.jutif.2024.5.4.2144.

[37] Ahmad Fajruddin Syauqi and D. D. Prasetya, “DEVELOPMENT OF HERBIFY APPLICATION WITH AI INTEGRATED UTILIZING YOLO V8 FOR OPTIMIZING HERBAL POTENTIAL IN INDONESIA,” *J. Tek. Inform. (JUTIF)*, vol. 5, no. 4, pp. 113-124, Jul. 2024. doi: 10.52436/1.jutif.2024.5.4.2094.

[38] Z. S. Hidayat, Y. A. . Wijaya, and R. Kurniawan, “OPTIMIZING YOLOV8 FOR AUTONOMOUS DRIVING: BATCH SIZE FOR BEST MEAN AVERAGE PRECISION (MAP),” *J. Tek. Inform. (JUTIF)*, vol. 5, no. 4, pp. 1147-1153, Jul. 2024. doi: 10.52436/1.jutif.2024.5.4.1626.

[39] N. Okila *et al.*, “Deep learning for accurate B-line detection and localization in lung ultrasound imaging,” *Frontiers in Artificial Intelligence*, 2025, doi: 10.3389/frai.2025.1560523.

[40] F. Conversano *et al.*, “Automatic approach for B-lines detection in lung ultrasound images using You Only Look Once algorithm,” *Journal of Ultrasound*, 2025, doi: 10.1007/s40477-025-01077-w.

[41] X. Wang *et al.*, “Enhanced pulmonary nodule detection with U-Net, YOLOv8, and Swin Transformer,” *BMC Medical Imaging*, vol. 25, p. 247, 2025, doi: 10.1186/s12880-025-01784-0.

[42] W. Zhu, X. Wang, J. Xing, X. S. Xu, and M. Yuan, “YOLOv8-BCD: a real-time deep learning framework for pulmonary nodule detection in computed tomography imaging,” *Quantitative Imaging in Medicine and Surgery*, vol. 15, no. 9, pp. 8189-8204, 2025, doi: 10.21037/qims-2025-824.

[43] K. Khotimah, S. Surono, and A. Thobirin, “Optimizing EfficientNet for imbalanced medical image classification using grey wolf optimization,” *Computer Science and Information Technologies*, vol. 6, no. 2, pp. 112–121, 2025, doi: 10.11591/csit.v6i2.pp112-121.

[44] H. Yang *et al.*, “Utilizing convolutional neural network and gray wolf optimization for image super-resolution,” *Journal of King Saud University - Science*, 2025, doi: 10.25259/JKSUS_162_2024.

[45] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, no. 1, pp. 159-174, 1977, doi: 10.2307/2529310.

[46] R. Kailasam and S. Balasubramanian, “Deep Learning for Pneumonia Detection: A Combined CNN and YOLO Approach,” *Hum.-Cent. Intell. Syst.*, vol. 5, pp. 44-62, 2025, doi: 10.1007/s44230-025-00091-9.

[47] M. R. Hasan, S. M. A. Ullah, and S. M. R. Islam, “Recent advancement of deep learning techniques for pneumonia prediction from chest X-ray image,” *Med. Rep.*, vol. 7, p. 100106, Oct. 2024, doi: 10.1016/j.hmedic.2024.100106.

[48] M. M. Kabir, M. F. Mridha, A. Rahman, M. A. Hamid, and M. M. Monowar, “Detection of COVID-19, pneumonia, and tuberculosis from radiographs using AI-driven knowledge distillation,” *Heliyon*, vol. 10, no. 5, p. e26801, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26801.

[49] J. M. Kimeu, M. Kisangiri, H. Mbelwa, and J. Leo, “Deep learning-based mobile application for the enhancement of pneumonia medical imaging analysis: A case-study of West-Meru Hospital,” *Informatics Med. Unlocked*, vol. 50, p. 101582, 2024, doi: 10.1016/j.imu.2024.101582.