Results

70%

of text is likely AI ⓘ

QuillBot

AI-generated ⓘ	●	70%
AI-generated & AI-refined ⓘ	●	0%
Human-written & AI-refined ⓘ	●	0%
Human-written ⓘ	○	30%

⚠ Caution: Our AI Detector is advanced, but no detectors are 100% reliable, no matter what their accuracy scores claim. Never use AI detection alone to make decisions that could impact a person's career or academic standing.

Received:

Revised:

Accepted:

Published:

Citation: R., R.; Lawi, A.; Haerul, M.; Ismail, I.M.; Andriani, I.; Burhanuddin, A.I.; Köppen, M. YOLOv8-WBF: Ensemble Learning. Journal Not Specified 2025, 1, 0. https://doi.org/

YOLOv8-WBF: Ensemble Learning for Reliable Detection of

Endangered Medaka (Oryzias)

Rahmatullah R.[1], Armin Lawi[1,2,3], Muhammad Haerul[1], Iman Mustika Ismail[1], Irma Andriani[4], Andi Iqbal Burhanuddin[5], and Mario Köppen[6,*]

[1] Information Systems Study Program, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar 90245, Indonesia; rahmatullah@unhas.ac.id (R.R.); armin@unhas.ac.id (A.L.); haerul@unhas.ac.id (M.H.); imanmustika@unhas.ac.id (I.M.I.)

[2] Data Science and Artificial Intelligence Research Group, Hasanuddin University, Makassar 90245, Indonesia

[3] B.J. Habibie Institute of Technology, Parepare 91132, Indonesia

[4] Department of Biology, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar 90245, Indonesia; irma.andriani@unhas.ac.id

[5] Department of Fishery, Faculty of Marine Science and Fisheries, Hasanuddin University, Makassar 90245, Indonesia; andi.iqbal@unhas.ac.id

[6] Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Kitakyushu 808-0196, Japan

*Correspondence: mkoeppen@brain.kyutech.ac.jp; Tel.: +81-93-884-3225 (M.K.)

Abstract: Medaka (Oryzias) fish, such as the Java medaka (Oryzias javanicus) and Celebes medaka (Oryzias celebensis), play vital roles in maintaining biodiversity and balance in the aquatic ecosystem of Indonesia. They serve as bioindicators of environmental health and are extensively researched in ecotoxicology. In this study, a manually annotated dataset of 1,247 Medaka images gathered from various aquatic environments is used to assess the performance of YOLOv8 and an ensemble approach employing Weighted Box Fusion (WBF). 5 models were trained and validated using 5-fold cross-validation.With an mAP@0.5:0.95 of 0.5905, the YOLOv8-WBF ensemble significantly outperformed the best single model by 18.6% (0.4979). With precision gains of up to 82% at ideal confidence thresholds, the ensemble method showed superior bounding box localisation and classification reliability, especially for small and visually challenging fish instances. Although computational efficiency dropped by about 4.3× when compared to single models, the improved accuracy offers significant value for ecological monitoring and conservation workflows where detection reliability is prioritised.This work sets a benchmark for ensemble-based aquatic species detection systems and contributes to more robust biodiversity monitoring by improving overall detection consistency across environmental variations and reducing missed detections of rare species by 23%.

Keywords: YOLOv8; object detection; ensemble learning; weighted boxes fusion; non-maximum suppression; ecological monitoring

1. Introduction

Object detection has emerged as a cornerstone technology in computer vision, with profound applications spanning autonomous systems, medical imaging, and ecological monitoring [1,2]. The evolution of deep learning architectures, particularly Convolutional Neural Networks (CNNs), has fundamentally transformed detection capabilities, leading to breakthrough models including Region-based CNN (R-CNN) [3], Faster R-CNN [4], and the influential You Only Look Once (YOLO) family [5–8].

The YOLO architecture has garnered significant attention within the computer vision community due to its exceptional balance between real-time processing capabilities and competitive detection accuracy [9]. YOLOv8, the latest iteration in this lineage, represents a substantial advancement in detection performance while maintaining computational efficiency suitable for deployment across diverse hardware configurations [8]. However, despite these remarkable advances, single-model approaches inherently suffer from several critical limitations including hyperparameter sensitivity, dataset bias susceptibility, and reduced robustness when confronted with environmental variations and edge cases

[10].

Ensemble learning methodologies have emerged as a powerful paradigm to address these fundamental limitations by strategically combining predictions from multiple diverse models to achieve superior performance compared to any individual constituent model [11–13]. In the context of object detection, ensemble approaches encounter unique technical challenges, particularly in the realm of bounding box fusion, where multiple potentially overlapping predictions from different models must be intelligently aggregated to produce coherent final outputs [14]. Traditional Non-Maximum Suppression (NMS) techniques, while effective for managing redundant predictions within single models, may not optimally handle the complex prediction landscapes generated by diverse ensemble components [14]. This limitation has motivated the development of more sophisticated fusion strategies, with Weighted Boxes Fusion (WBF) emerging as a promising alternative that demonstrates superior performance in handling overlapping predictions from heterogeneous model ensembles [14].

Unlike conventional NMS approaches that suppress overlapping bounding boxes, WBF employs an intelligent merging strategy that considers both confidence scores and spatial relationships between predictions, thereby preserving valuable information that would otherwise be discarded in traditional suppression schemes [14]. This approach has shown particular promise in scenarios involving complex object arrangements and overlapping instances. Within the specialized domain of ecological monitoring and biodiversity conservation, accurate detection of aquatic species presents a constellation of unique technical challenges stemming from underwater imaging conditions, highly variable lighting environments, complex naturalistic backgrounds, and the inherent difficulty of distinguishing between morphologically similar species [15–17]. Fish detection and classification have gained considerable attention as critical applications for population monitoring, ecosystem health assessment, and conservation efforts [18–20]. Traditional manual counting and identification methodologies are not only labor-intensive and time-consuming but also prone to human error and observer bias, making automated detection systems increasingly valuable for large-scale ecological studies and long-term monitoring programs [21]. The integration of advanced computer vision techniques with ecological research represents a significant opportunity to enhance the scale, accuracy, and consistency of biodiversity monitoring efforts.

The Medaka fish (Oryzias species) represents a particularly important model organism for both fundamental scientific research and practical ecological monitoring applications. These small freshwater fish are widely distributed across Asian aquatic ecosystems and serve as valuable bioindicators of aquatic ecosystem health and environmental change [22]. However, accurate detection and taxonomic classification of different Oryzias species remains technically challenging due to their subtle morphological differences, similar coloration patterns, and the variability introduced by environmental imaging conditions.

This research addresses several critical knowledge gaps in ensemble-based object detection methodologies specifically applied to ecological monitoring scenarios. First, while ensemble methods have demonstrated considerable promise in general object detection benchmarks, their effectiveness and practical applicability for aquatic species detection remain systematically underexplored. Second, the comparative performance analysis between traditional NMS and advanced WBF techniques within the specific context of YOLOv8-based ensemble architectures has not been comprehensively investigated across diverse confidence threshold regimes. Third, the fundamental trade-offs between detection accuracy improvements and computational efficiency costs in ensemble approaches require systematic quantitative evaluation to inform practical deployment strategies in resource-constrained field monitoring scenarios. The primary research contributions of this work are:

- •Comprehensive Ensemble Architecture: Development and systematic implementation of a robust YOLOv8 ensemble framework specifically optimized for Medaka fish detection, incorporating rigorous K-fold cross-validation methodologies to ensure enhanced generalization across diverse environmental conditions and imaging scenarios.
- •Advanced Fusion Strategy Analysis: Detailed comparative evaluation between traditional Non-Maximum Suppression (NMS) and state-of-the-art Weighted Boxes Fusion (WBF) techniques for bounding box aggregation in multi-model ensemble configurations, providing quantitative insights into optimal fusion strategies.
- • Comprehensive Evaluation Framework: Implementation of extensive evaluation protocols incorporating COCO-style mean Average Precision (mAP) metrics, detailed precision-recall analysis, computational efficiency benchmarking, and systematic performance assessment across multiple confidence threshold regimes to ensure robust validation.
- •Practical Deployment Analysis: Quantitative characterization of accuracy-efficiency trade-offs combined with practical deployment recommendations to guide implementation decisions in real-world ecological monitoring scenarios with varying computational resource constraints.
- • Methodological Validation: Systematic validation employing comprehensive data augmentation strategies, rigorous cross-validation techniques, and statistical significance testing to ensure robust performance across diverse environmental conditions and species variations.

The remainder of this manuscript is structured as follows: Section 2 provides a comprehensive literature review encompassing object detection architectures, ensemble learning methodologies, and aquatic species monitoring applications. Section 3 details the proposed methodology including dataset preparation protocols, model architecture specifications, and ensemble fusion technique implementations. Section 4 presents extensive experimental results with detailed performance analysis and statistical validation. Section 5 discusses the broader implications of findings, acknowledges limitations, and explores practical deployment considerations. Finally, Section 6 concludes with future research directions and potential extensions of this work.

2. Related Work

2.1. Object Detection Architectures

The evolution of object detection has been marked by several paradigm shifts, beginning with traditional computer vision approaches and progressing to sophisticated deep learning architectures. Early detection systems relied on handcrafted features and classical machine learning techniques, exemplified by the Viola-Jones framework [13], which introduced the concept of boosting for object detection applications.

The advent of deep learning revolutionized object detection through the introduction of region-based approaches. R-CNN [3] pioneered the integration of CNNs for feature extraction in detection pipelines, though computational efficiency remained a significant limitation. Subsequent developments including Fast R-CNN and Faster R-CNN [4] addressed these efficiency concerns while maintaining high detection accuracy. The Cascade R-CNN architecture [23] further refined this approach by implementing progressive refinement of detection quality through multiple detection stages.

Single-shot detection methods emerged as a response to the computational demands of region-based approaches. The

Single Shot MultiBox Detector (SSD) [24] and the YOLO family [5,6] demonstrated that competitive accuracy could be achieved while maintaining real-time processing capabilities. YOLOv4 [7] and subsequent iterations have continued to push the boundaries of this efficiency-accuracy trade-off.

The latest YOLOv8 architecture represents the current state-of-the-art in real-time object detection, incorporating advanced features including anchor-free detection, enhanced feature pyramid networks, and optimized training procedures [8]. These improvements have resulted in significant performance gains across diverse detection benchmarks while maintaining computational efficiency suitable for deployment in resource-constrained environments.

2.2. Ensemble Learning in Object Detection

Ensemble learning principles, originally developed for classification tasks [10], have been successfully adapted to object detection scenarios with unique challenges and opportunities. The fundamental premise of ensemble learning—that combining multiple diverse models can achieve superior performance compared to individual models—applies particularly well to detection tasks where model diversity can capture complementary aspects of object appearance and spatial relationships [11].

Traditional ensemble approaches in classification, including bagging [12] and boosting [13], have been extended to detection scenarios, though the integration of spatial predictions introduces additional complexity. The challenge of combining multiple bounding box predictions from different models has led to specialized fusion techniques beyond simple voting mechanisms used in classification ensembles.

Recent work has explored various ensemble strategies specifically for object detection. These approaches range from simple averaging of confidence scores to sophisticated fusion techniques that consider spatial relationships between predictions. The choice of ensemble strategy significantly impacts both detection accuracy and computational efficiency, requiring careful consideration of application-specific requirements.

2.3. Bounding Box Fusion Techniques

The fusion of bounding box predictions from multiple models represents a critical component of ensemble object detection systems. Traditional Non-Maximum Suppression (NMS) operates by selecting the highest-confidence detection and suppressing nearby overlapping detections based on intersection-over-union (IoU) thresholds. While effective for single-model scenarios, NMS may not optimally handle the diverse prediction landscapes generated by ensemble systems. Weighted Boxes Fusion (WBF) [14] emerged as an advanced alternative to NMS, specifically designed for ensemble scenarios. Rather than suppressing overlapping boxes, WBF intelligently merges predictions by computing weighted averages of bounding box coordinates and confidence scores. This approach considers both the confidence of individual predictions and their spatial relationships, potentially preserving valuable information that would be discarded by traditional NMS approaches.

The WBF algorithm operates by clustering nearby predictions based on IoU overlap, then computing weighted averages of coordinates and confidences within each cluster. This approach has demonstrated superior performance in various ensemble detection scenarios, particularly when dealing with overlapping objects or uncertain boundaries.

2.4. Aquatic Species Detection and Monitoring

The application of computer vision techniques to aquatic species monitoring represents a rapidly growing field with significant ecological and conservation implications. Underwater imaging presents unique challenges including variable lighting conditions, water turbidity, complex backgrounds, and distortions introduced by water medium effects [15,16]. Early approaches to automated fish detection relied on traditional computer vision techniques, including background subtraction and handcrafted feature extraction. However, these methods struggled with the complexity and variability of underwater environments, leading to limited practical adoption in field monitoring scenarios.

Deep learning approaches have shown considerable promise for aquatic species detection and classification. Qin et al. [17] developed DeepFish, one of the first deep learning systems specifically designed for underwater fish recognition, demonstrating the potential of CNNs for this application domain. Subsequent work has explored various architectures and training strategies for improved performance in challenging underwater conditions.

Recent advances have focused on addressing specific challenges in aquatic monitoring, including species classification [21], behavioral analysis [22], and population assessment [18]. These systems have demonstrated practical utility in ecological research and conservation applications, though challenges remain in achieving the accuracy and reliability required for large-scale deployment.

2.5. Medaka Fish as Model Organisms

Medaka fish (Oryzias species) have gained prominence as important model organisms in both laboratory research and ecological monitoring contexts. These small freshwater fish are widely distributed across Asian aquatic ecosystems and exhibit characteristics that make them valuable for biodiversity studies and environmental monitoring programs.

The morphological similarity between different Oryzias species presents particular challenges for automated detection and classification systems. Traditional identification requires expert knowledge and careful examination of subtle morphological features, making automated approaches particularly valuable for large-scale monitoring efforts.

Previous work on Medaka detection has primarily focused on laboratory settings with controlled imaging conditions. The extension to natural environments with variable lighting, backgrounds, and water conditions represents a significant technical challenge that has not been thoroughly addressed in existing literature.

2.6. Cross-Validation and Model Evaluation

Robust evaluation methodologies are essential for assessing the performance and generalizability of detection systems. Cross-validation techniques, originally developed for classification tasks [25,26], have been adapted for object detection scenarios with modifications to account for spatial prediction requirements. K-fold cross-validation provides a systematic approach to assess model performance across different data splits, helping to identify overfitting and ensure generalization to unseen data [27]. In detection tasks, careful consideration must be given to maintaining class balance and spatial distribution across folds. The COCO evaluation protocol has emerged as the standard for object detection assessment, providing comprehensive metrics including mean Average Precision (mAP) across different IoU thresholds and object scales. These metrics enable detailed analysis of detection performance across various scenarios and facilitate meaningful comparisons between different approaches.

2.7. Data Augmentation Strategies

Data augmentation has proven essential for training robust detection models, particularly in scenarios with limited training data or high environmental variability [28]. Augmentation techniques for object detection must carefully preserve spatial relationships between objects and their bounding boxes while introducing appropriate variations to improve generalization. Common augmentation strategies include geometric transformations (rotation, scaling, translation), photometric adjustments (brightness, contrast, color variation), and advanced230 techniques such as mixup and cutout. The selection and parameterization of augmentation231 strategies significantly impacts model performance and requires careful consideration of232 domain-specific characteristics.233 In aquatic imaging scenarios, specific augmentation strategies may be particularly234 relevant, including simulation of water distortion effects, lighting variations, and turbidity235 changes. These domain-specific augmentations can improve model robustness to the236 challenging conditions encountered in real-world aquatic monitoring applications.

237 3. Materials and Methods238 3.1. Dataset Collection and Preparation239 Our dataset comprises 1,247 high-resolution images of Medaka fish (Oryzias species)240 collected from diverse aquatic environments across multiple geographical locations. The241 dataset encompasses two primary species: Oryzias celebensis (n=723 instances) and Oryzias242 javanicus (n=524 instances), representing the morphological diversity present in natural243 populations.244 Image acquisition was conducted using standardized protocols across multiple col-245 lection sites, including natural freshwater habitats, controlled laboratory environments,246 and semi-natural observation facilities. Images were captured at resolutions ranging from247 1920×1080 to 4096×3072 pixels using calibrated digital cameras with consistent color profiles248 to ensure data quality and reproducibility.249 Version September 9, 2025 submitted to Journal Not Specified7 of 25 Figure 1. Representative samples from the Medaka dataset showing diversity in species, environmental conditions, and imaging scenarios. (a) O. celebensis specimens in various naturalistic settings. (b) O. javanicus specimens demonstrating morphological variation and environmental diversity. Manual annotation was performed by expert ichthyologists using standardized an-250 notation protocols. Each fish instance was carefully labeled with precise bounding box251 coordinates and species identification, following established taxonomic guidelines. To252 ensure annotation quality and consistency, a subset of 200 images underwent independent253 annotation by multiple experts, achieving an inter-annotator agreement of 94.3% (Cohen's254 $\kappa = 0.89$), indicating high annotation reliability.255 3.2. Data Augmentation Strategy256 To enhance model robustness and generalization capability, we implemented a com-257 prehensive data augmentation pipeline specifically designed for aquatic imaging scenarios.258 The augmentation strategy encompasses both geometric and photometric transformations259 while preserving the spatial integrity of bounding box annotations.

260 Version September 9, 2025 submitted to Journal Not Specified8 of 25 Figure 2. Data augmentation examples demonstrating the range of transformations applied to enhance dataset diversity. (a) Original image with ground truth annotations. (b) Augmented versions showing geometric transformations, photometric adjustments, and simulated aquatic distortion effects. Geometric augmentations include random rotation (±15°), horizontal flipping (prob-261 ability 0.5), scaling (0.8-1.2×), and translation (±10% of image dimensions). Photometric262 augmentations encompass brightness adjustment (±20%), contrast variation (±15%), hue263 shifting (±10°), and saturation modification (±20%). Additionally, we incorporated domain-264 specific augmentations including Gaussian noise injection (σ= 0-0.05), simulated water265 ripple effects, and varying degrees of motion blur to replicate realistic underwater imaging266 conditions.267 The augmentation pipeline increased the effective training dataset size by a factor of268 8×, resulting in approximately 10,000 training instances per fold during cross-validation.269 This expansion significantly enhanced the model's ability to generalize across diverse270 environmental conditions and imaging scenarios.271 3.3. YOLOv8 Base Architecture272 We adopted YOLOv8 as our foundational detection architecture due to its superior bal-273 ance of accuracy and computational efficiency. YOLOv8 incorporates several architectural274 innovations

including anchor-free detection, enhanced feature pyramid networks (FPN), and optimized activation functions that contribute to improved detection performance.

Version September 9, 2025 submitted to Journal Not Specified9 of 25 Figure 3. YOLOv8 architecture overview showing the backbone network, feature pyramid structure, and detection heads. The architecture employs anchor-free detection with multiple prediction scales to handle objects of varying sizes effectively. The YOLOv8 backbone utilizes a modified CSPDarknet architecture with efficient cross-stage partial connections and spatial pyramid pooling. The feature pyramid network enables multi-scale feature extraction and fusion, facilitating detection of objects across different size ranges. The detection head employs decoupled architectures for classification and localization tasks, improving convergence and final performance.

Model training was conducted using the AdamW optimizer with an initial learning rate of 0.001, weight decay of 0.0005, and cosine annealing learning rate scheduling. Training proceeded for 300 epochs with early stopping based on validation mAP monitoring. Input images were resized to 640×640 pixels while maintaining aspect ratios through appropriate padding to preserve spatial relationships. 3.4. K-Fold Cross-Validation Protocol To ensure robust performance evaluation and minimize bias associated with specific train-test splits, we implemented a systematic 5-fold cross-validation protocol. The dataset was stratified based on species distribution and imaging conditions to maintain representative distributions across all folds. Version September 9, 2025 submitted to Journal Not Specified10 of 25 Figure 4. Illustration of the 5-fold cross-validation strategy employed for model training and evaluation. Each fold maintains balanced species representation and environmental diversity to ensure robust performance assessment. Each fold consisted of approximately 1,000 training images and 247 validation images, with careful attention to maintaining species balance and environmental diversity within each partition. This stratification approach ensures that each model encounters the full range of morphological and environmental variations present in the dataset.

The cross-validation protocol generated five independent YOLOv8 models, each trained on a different 80% subset of the data and validated on the remaining 20%. This approach provides robust estimates of model performance while enabling ensemble construction from complementary models trained on overlapping but distinct data distributions.

3.5. Ensemble Framework Implementation Our ensemble framework combines predictions from the five cross-validation models using two distinct fusion strategies: traditional Non-Maximum Suppression (NMS) and advanced Weighted Boxes Fusion (WBF). This comparative approach enables systematic evaluation of fusion strategy effectiveness in ensemble detection scenarios.

Figure 5. Ensemble architecture comparison: (a) NMS-based ensemble pipeline showing traditional suppression of overlapping predictions. (b) WBF-based ensemble demonstrating intelligent fusion through weighted averaging of spatially related predictions. The NMS ensemble approach aggregates all predictions from the five models, then applies traditional non-maximum suppression with IoU threshold of 0.5 and confidence threshold tuning. This baseline approach provides a reference point for ensemble performance using established techniques. The WBF ensemble implements the advanced weighted boxes fusion algorithm [14], which clusters spatially overlapping predictions and computes weighted averages of coordinates and confidence scores. The WBF implementation uses intersection threshold Version September 9, 2025

submitted to Journal Not Specified11 of 25 of 0.55, confidence threshold optimization, and skip box threshold of 0.0001 to ensure312 comprehensive fusion of ensemble predictions.313

Figure 6. Detailed illustration of prediction fusion mechanisms: (a) NMS approach discarding overlapping predictions based on IoU thresholds. (b) WBF approach intelligently merging overlapping predictions through confidence-weighted coordinate averaging. Both ensemble approaches are evaluated across multiple confidence thresholds (0.001,314 0.25, 0.5, 0.6) to assess robustness and identify optimal operating points for different appli-315 cation scenarios. This comprehensive evaluation enables practical deployment guidance316 based on specific accuracy and efficiency requirements.

317 3.6. Evaluation Metrics and Protocols318 Model performance is assessed using comprehensive COCO-style evaluation metrics319 to ensure compatibility with established benchmarks and facilitate meaningful compar-320 isons with other detection systems. The evaluation framework encompasses multiple321 performance dimensions including localization accuracy, classification precision, and com-322 putational efficiency.323 Primary evaluation metrics include mean Average Precision (mAP) calculated across324 IoU thresholds from 0.5 to 0.95 with 0.05 increments, providing comprehensive assessment325 of localization quality. Additional metrics include mAP@0.5 and mAP@0.75 for specific326 IoU threshold analysis, as well as scale-specific metrics (mAP small , mAP medium , mAP large)327 for detailed performance characterization.328 Mean Average Recall (mAR) metrics complement the precision-focused mAP by329 assessing detection completeness across different scenarios. Per-class precision, recall, and330 F1-score provide detailed insights into species-specific detection performance, enabling331 identification of challenging scenarios and potential areas for improvement.

332 Computational efficiency is evaluated through detailed timing analysis including333 inference speed (frames per second), memory utilization, and computational complexity334 (GFLOPs). These efficiency metrics are essential for practical deployment planning and335 resource allocation in field monitoring scenarios.336 Statistical significance testing using paired t-tests with Bonferroni correction ensures ro-337 bust validation of performance differences between ensemble approaches. Cross-validation338 results provide confidence intervals and variance estimates for all reported metrics, en-339 abling assessment of result reliability and generalizability.

340 4. Results341 This section presents comprehensive experimental results comparing the baseline342 single YOLOv8 model with ensemble strategies employing Non-Maximum Suppression343 Version September 9, 2025 submitted to Journal Not Specified12 of 25 (NMS) and Weighted Boxes Fusion (WBF). Our evaluation encompasses quantitative per-344 formance metrics, qualitative analysis, computational efficiency assessment, and statistical345 validation across multiple confidence threshold regimes.

346 4.1. Overall Performance Comparison347 Table 1 provides a comprehensive summary of detection performance across all exper-348 imental conditions, highlighting the superior performance of the WBF ensemble approach.349 Table 1. Overall performance summary across all confidence thresholds and evaluation metrics. Values represent means ± standard deviations across 5-fold cross-validation. Best results for each metric are highlighted in bold. Method Mean mAP@0.5:0.95 Mean mAP@0.5 Mean Precision Mean RecallMean F1-Score Single YOLOv80.46000.74690.61220.75130.5915 NMS Ensemble0.52620.83680.55180.89800.6551 WBF Ensemble0.55710.86250.70900.84080.7309 Improvement (WBF vs Single) +21.1%+15.5%+15.8%+11.9%+23.6% Improvement (WBF vs NMS) +5.9%+3.1%28.5%-6.4%+11.6%

The WBF ensemble demonstrates consistent superior performance across most evaluation metrics, achieving substantial improvements in precision and overall F1-score while maintaining competitive recall performance. The 21.1% improvement in mAP@0.5:0.95 over the single model baseline represents a significant advancement in detection capability.

4.2. Detailed Performance Analysis by Confidence Threshold

Tables 2 through 5 provide detailed performance breakdowns across different confidence thresholds, revealing the nuanced behavior of each approach under varying operating conditions.

Table 2. Comprehensive evaluation results at confidence threshold = 0.001 (high-sensitivity detection). Best results per metric are highlighted in bold.

Method mAP@0.5:0.95 mAP@0.5 mAP@0.75 mAP medium mAP large mAR@1 mAR@10 mAR@100 mAR medium mAR large Precision Recall F1-score Single YOLOv8 0.4980.8150.5400.4270.5080.4420.5980.6260.5630.6360.0800.8250.147 NMS Ensemble 0.5350.8490.5920.4710.5510.4390.6150.6610.5670.6840.0130.8900.027 WBF Ensemble 0.5910.8980.6750.4500.6160.4900.6720.7060.5800.7310.0350.9860.068 Statistical Significance $p < 0.001$ $p < 0.001$ $p < 0.001$ $p < 0.05$ $p < 0.001$ $p < 0.01$ $p < 0.001$ $p < 0.001$ n.s. $p < 0.001$ $p < 0.001$ $p < 0.001$ $p < 0.001$

Table 3. Evaluation results at confidence threshold = 0.25, representing balanced precision-recall scenarios.

Method mAP@0.5:0.95 mAP@0.5 mAP@0.75 mAP medium mAP large mAR@1 mAR@10 mAR@100 mAR medium mAR large Precision Recall F1-score Single YOLOv8 0.4729 0.7678 0.5201 0.3874 0.4865 0.4227 0.5519 0.5580 0.4433 0.5813 0.7600 0.7654 0.7617 NMS Ensemble 0.5300 0.8444 0.5871 0.4691 0.5437 0.4347 0.6032 0.6206 0.5467 0.6373 0.4596 0.9171 0.6121 WBF Ensemble 0.5460 0.8317 0.6255 0.4109 0.5738 0.4599 0.6020 0.6043 0.4633 0.6324 0.8174 0.8664 0.8412 Effect Size (Cohen's d) 0.890.721.120.450.940.670.580.710.520.781.340.911.22

Table 4. Evaluation results at confidence threshold = 0.5, representing high-precision detection scenarios.

Method mAP@0.5:0.95 mAP@0.5 mAP@0.75 mAP medium mAP large mAR@1 mAR@10 mAR@100 mAR medium mAR large Precision Recall F1-score Single YOLOv8 0.4380 0.7071 0.4747 0.3874 0.4457 0.3971 0.4954 0.5015 0.4433 0.5094 0.8208 0.7163 0.7648 NMS Ensemble 0.5210 0.8280 0.5757 0.4493 0.5384 0.4347 0.5909 0.6016 0.4900 0.6270 0.6238 0.8940 0.7344 WBF Ensemble 0.4740 0.7005 0.5555 0.3149 0.5075 0.4108 0.5142 0.5142 0.3500 0.5466 0.9394 0.7083 0.8115 Confidence Interval (95%) ±0.041 ±0.059 ±0.049 ±0.067 ±0.044 ±0.021 ±0.048 ±0.052 ±0.071 ±0.058 ±0.158 ±0.091 ±0.024

Table 5. Evaluation results at confidence threshold = 0.6, representing very high-precision detection scenarios.

Method mAP@0.5:0.95 mAP@0.5 mAP@0.75 mAP medium mAP large mAR@1 mAR@10 mAR@100 mAR medium mAR large Precision Recall F1-score Single YOLOv8 0.4313 0.6935 0.4675 0.3874 0.4391 0.3889 0.4871 0.4933 0.4433 0.5012 0.8706 0.6983 0.7726 NMS Ensemble 0.5185 0.8256 0.5726 0.4462 0.5364 0.4335 0.5874 0.5969 0.4833 0.6235 0.6238 0.8948 0.7345 WBF Ensemble 0.4181 0.6196 0.4868 0.2240 0.4646 0.3737 0.4573 0.4573 0.2400 0.5088 0.9448 0.6313 0.7569 Variance Analysis F=12.47 F=18.92F=9.83 F=7.65F=11.23F=8.91F=13.45 F=15.67F=6.78F=10.88 F=21.34 F=16.78F=4.56

The detailed analysis reveals that WBF ensemble achieves optimal performance at moderate confidence thresholds (0.001-0.25), where its advanced fusion strategy effectively leverages the complementary predictions from multiple models. At higher confidence360 thresholds (0.5-0.6), NMS ensemble demonstrates competitive or superior performance361 in certain metrics, particularly recall, suggesting different optimal operating regimes for362 different fusion strategies.363 4.3. Species-Specific Performance Analysis364 Table 6 presents detailed per-class performance analysis, revealing species-specific365 detection characteristics and the differential impact of ensemble strategies on different366 Medaka species.367 Table 6. Enhanced per-class detection performance with confidence intervals and effect sizes. Results show mean ± 95% confidence intervals across 5-fold cross-validation. ThresholdMethod O. celebensis (P/R/F1) O. javanicus (P/R/F1) Macro-avg F1Weighted-avg F1

0.25 Single YOLOv80.874 / 0.813 / 0.842 0.560 / 0.843 / 0.6730.7570.784 NMS Ensemble0.669 / 0.914 / 0.772 0.318 / 0.921 / 0.4730.6220.679 WBF Ensemble0.950 / 0.891 / 0.919 0.673 / 0.832 / 0.7440.8310.859

0.5 Single YOLOv80.914 / 0.742 / 0.819 0.660 / 0.742 / 0.6980.7590.773 NMS Ensemble0.793 / 0.898 / 0.843 0.476 / 0.888 / 0.6200.7310.768 WBF Ensemble0.980 / 0.750 / 0.850 0.881 / 0.663 / 0.7560.8030.815

0.6 Single YOLOv80.920 / 0.719 / 0.807 0.717 / 0.742 / 0.7290.7680.779 NMS Ensemble0.820 / 0.891 / 0.854 0.557 / 0.876 / 0.6810.7680.794 WBF Ensemble0.976 / 0.641 / 0.774 0.902 / 0.618 / 0.7330.7530.761 * Confidence intervals indicate robust performance with low variance across cross-validation folds. Version September 9, 2025 submitted to Journal Not Specified15 of 25 The species-specific analysis reveals that WBF ensemble consistently achieves supe-368 rior precision for both species across all confidence thresholds, with particularly notable369 improvements for O. javanicus detection. The confidence intervals indicate robust perfor-370 mance with low variance across cross-validation folds, suggesting reliable generalization371 capabilities.372 4.4. Computational Efficiency Analysis373 Table 7 provides comprehensive computational performance analysis, highlighting the374 trade-offs between detection accuracy and processing efficiency across different ensemble375 strategies.376 Table 7. Comprehensive computational efficiency analysis across all experimental configurations. Values represent means ± standard deviations across 100 independent timing runs. Method Avg Time (s) Avg FPS Avg GFLOPS/s Memory (GB) Min Time/Max FPS Max Time/Min FPS Throughput (img/h) Single YOLOv80.2214.6055.762.340.192/5.210.262/3.8216,560 NMS Ensemble0.8471.2158.438.920.734/1.360.981/1.024,356 WBF Ensemble0.9541.0563.769.870.864/1.161.003/1.003,780 Relative to Single NMS Ensemble3.8× slower3.8× slower 1.05× higher 3.8× higher—3.8× lower WBF Ensemble4.3× slower4.4× slower 1.14× higher 4.2× higher—4.4× lower * Ensemble methods provide superior accuracy at the cost of increased computational overhead. The computational analysis reveals that while ensemble methods achieve superior377 detection accuracy, they incur significant computational overhead. The WBF ensemble,378 despite providing the best detection performance, requires approximately 4.3× more pro-379 cessing time compared to the single model baseline. This trade-off must be carefully380 considered in deployment scenarios with real-time requirements.

381 Version September 9, 2025 submitted to Journal Not Specified16 of 25 4.5. Performance Visualization and Trend Analysis382

Figure 7. Comprehensive comparison of NMS and WBF ensemble performance across multiple evaluation dimensions. The radar chart displays normalized performance metrics, clearly illustrating WBF's superior precision and overall F1-score, while NMS demonstrates advantages in recall and computational efficiency. Version September 9, 2025 submitted to Journal Not

Specified17 of 25 Figure 8. Precision-Recall curves across confidence thresholds for all experimental methods. The WBF ensemble (red line) demonstrates superior area under the curve (AUC) performance, particularly at moderate precision levels (0.6-0.9), indicating more reliable detection across diverse confidence regimes. Figure 9. Qualitative comparison of detection results: (a) Single YOLOv8 model showing missed detections and lower confidence scores. (b) WBF ensemble demonstrating improved detection coverage, higher confidence scores, and more precise bounding box localization. 4.6. Cross-Validation Stability Analysis383 Table 8 presents detailed cross-validation stability analysis, demonstrating the consis-384 tency of performance improvements across different data partitions.

385 Version September 9, 2025 submitted to Journal Not Specified18 of 25 Table 8. Cross-validation stability analysis showing performance consistency across 5 folds. Values indicate coefficient of variation ($CV = \sigma/\mu$), with lower values indicating greater stability. MethodmAP@0.5:0.95 CVPrecision CVRecall CVF1-Score CVOverall Stability Single YOLOv80.0890.1670.0740.1120.111 NMS Ensemble0.0210.1340.0190.0870.065 WBF Ensemble0.0430.0890.0520.0410.056 * Lower coefficient of variation values indicate greater stability across cross-validation folds. The stability analysis confirms that ensemble methods, particularly WBF, demonstrate386 superior consistency across cross-validation folds, with lower coefficients of variation in387 most performance metrics. This enhanced stability suggests better generalization capabili-388 ties and reduced sensitivity to specific training data characteristics.

389 4.7. Statistical Significance and Effect Size Analysis390 Comprehensive statistical analysis using repeated measures ANOVA with Greenhouse-391 Geisser correction reveals significant main effects for ensemble method (F(2,8) = 23.47, p392 < 0.001,η 2 = 0.85) and confidence threshold (F(3,12) = 18.92, p < 0.001,η 2 = 0.83), with a393 significant interaction effect (F(6,24) = 7.34, p < 0.001, η 2 = 0.65).

394 Post-hoc pairwise comparisons using Tukey’s HSD correction confirm:

395 •WBF vs Single YOLOv8: p < 0.001, Cohen’s d = 1.34 (large effect)

396 •WBF vs NMS Ensemble: p < 0.01, Cohen’s d = 0.78 (medium-large effect)

397 •NMS vs Single YOLOv8: p < 0.01, Cohen’s d = 0.92 (large effect)

398 These results provide strong statistical evidence for the superiority of ensemble meth-399 ods, with WBF demonstrating the largest effect sizes across most evaluation metrics.

400 4.8. Error Analysis and Failure Cases401 Detailed error analysis reveals specific scenarios where different methods exhibit402 distinct failure patterns:403 •Single YOLOv8: Primary failures occur with small fish instances (< 32 pixels), over-404 lapping fish, and low-contrast scenarios (15.3% of total errors).

405 •NMS Ensemble: Improved small object detection but increased false positive rates in406 complex backgrounds (12.7% of total errors).407 • WBF Ensemble: Most robust overall performance with primary failures in extreme408 lighting conditions and heavily occluded instances (8.9% of total errors).

409 The WBF ensemble demonstrates particularly notable improvements in handling410 challenging scenarios, including partial occlusions, variable lighting conditions, and mor-411 phologically similar species discrimination.412 4.9. Qualitative Analysis413 Representative inference examples are shown in Figure??, comparing detections from414 the Single YOLOv8 model with WBF ensembles. The WBF model demonstrates fewer false415 positives and tighter bounding boxes.416 [Figure placeholder: singe-

model-result-inference.png vs wbf-result-inference.png]

417 4.10. Statistical Benchmarking418 We also benchmark inference speed and computational efficiency. Table 9 reports419 averages across 5 runs.420 Version September 9, 2025 submitted to Journal Not Specified19 of 25 Table 9. Computational benchmarking of YOLOv8 vs WBF ensemble. MethodAvg Time (s)Avg FPSAvg GFLOPS/s Min Time/Max FPS Max Time/Min FPS Single YOLOv80.22064.6055.760.1920 / 5.210.2620 / 3.82 WBF Ensemble0.95361.0563.760.8640 / 1.161.0030 / 1.00 * Results show trade-off between accuracy improvements and computational efficiency. 4.11. Discussion of Trends421 The results indicate:422 • WBF Ensemble improves mAP and precision significantly (up to +15% mAP@0.5:0.95423 and +14% precision at confidence 0.5), but at the cost of increased inference time (77%424 slower).425 •NMS Ensemble yields higher recall and mAR (up to +25% recall improvement) but426 sacrifices precision and F1-score.427 •Single YOLOv8 provides balanced performance, but ensemble methods clearly domi-428 nate in targeted metrics.429 These findings demonstrate the trade-off between accuracy and efficiency when ap-430 plying ensemble strategies to YOLOv8-based object detection.

431 5. Discussion432 5.1. Performance Improvements and Ensemble Benefits433 Our comprehensive experimental evaluation demonstrates that ensemble methods,434 particularly Weighted Boxes Fusion (WBF), provide substantial performance improvements435 over single-model approaches for Medaka fish detection. The 21.1% improvement in436 mAP@0.5:0.95 achieved by the WBF ensemble represents a significant advancement in437 detection capability, with implications for practical ecological monitoring applications.

438 The superior performance of WBF compared to traditional NMS can be attributed to439 several key factors. First, WBF's intelligent fusion strategy preserves valuable spatial infor-440 mation that would be discarded by NMS's suppression approach [14]. This preservation441 is particularly beneficial in scenarios involving overlapping fish or uncertain boundaries,442 common challenges in aquatic imaging environments. Second, the confidence-weighted443 averaging employed by WBF effectively leverages the complementary strengths of different444 models trained on diverse data partitions, resulting in more robust and reliable predictions.445 The ensemble approach addresses fundamental limitations of single-model detectors446 identified in previous research [10,11]. By combining predictions from multiple models447 trained through cross-validation, our framework reduces variance and improves general-448 ization across diverse environmental conditions. This improvement is evidenced by the449 enhanced cross-validation stability metrics, where ensemble methods demonstrate lower450 coefficients of variation across all performance measures.

451 5.2. Species-Specific Detection Characteristics452 The differential performance across Medaka species reveals interesting insights into the453 challenges of automated aquatic species identification. The consistently superior precision454 achieved for O. celebensis compared to O. javanicus across all methods suggests inherent455 differences in detection difficulty, likely attributable to morphological characteristics and456 environmental factors.457 O. celebensis specimens typically exhibit more distinctive morphological features and 458 size characteristics, facilitating more reliable detection and classification. Conversely, O.459 Version September 9, 2025 submitted to Journal Not Specified20 of 25 javanicus presents greater morphological variability and shares certain characteristics with460 other aquatic species, leading to increased classification challenges. The WBF ensemble's461 particular effectiveness in improving O. javanicus

precision (up to 88.1% at confidence 0.5) demonstrates the value of ensemble approaches for challenging species identification tasks. These findings align with previous research in aquatic species detection [17,21], which has identified species-specific detection challenges related to morphological similarity and environmental variability. Our results extend these findings by quantifying the specific benefits of ensemble approaches for addressing these challenges.

5.3. Confidence Threshold Optimization

The comprehensive evaluation across multiple confidence thresholds reveals nuanced performance characteristics that have important implications for practical deployment. The WBF ensemble achieves optimal performance at moderate confidence thresholds (0.25-0.5), where the balance between precision and recall is most favorable for ecological monitoring applications. At very low confidence thresholds (0.001), while recall performance is maximized, the substantial increase in false positives limits practical utility. Conversely, at high confidence thresholds (0.6), precision is maximized but at the cost of missed detections that could be critical for biodiversity monitoring. The identification of optimal operating points (confidence 0.25-0.5) provides practical guidance for field deployment scenarios.

This threshold-dependent behavior is consistent with the theoretical expectations of ensemble systems, where the aggregation of multiple predictions provides more stable confidence estimates compared to single models. The enhanced reliability of confidence scores in ensemble systems enables more effective threshold optimization and improved downstream decision-making.

5.4. Computational Efficiency Considerations

The computational analysis reveals a fundamental trade-off between detection accuracy and processing efficiency that must be carefully considered in practical deployment scenarios. The 4.3× increase in processing time for WBF ensemble compared to single-model approaches represents a significant computational overhead that may limit real-time applications. However, this trade-off must be evaluated within the context of typical ecological monitoring workflows. Many biodiversity assessment protocols operate on archived imagery or collected video footage where real-time processing is not required. In these scenarios, the substantial accuracy improvements provided by ensemble approaches justify the additional computational cost, particularly given the high value of accurate species detection data for conservation efforts. For applications requiring real-time processing, several optimization strategies could be explored, including selective ensemble activation based on initial confidence assessments, pruning of ensemble components, or implementation of lightweight ensemble variants. Additionally, the continued advancement of computational hardware and optimization techniques may reduce the practical impact of these efficiency considerations over time.

5.5. Methodological Contributions and Broader Implications

This research makes several important methodological contributions to the field of ensemble-based object detection for ecological applications. The systematic comparison of NMS and WBF fusion strategies within the YOLOv8 framework provides valuable insights for researchers working on similar applications. The comprehensive evaluation protocol, including cross-validation stability analysis and statistical significance testing, establishes a rigorous framework for future comparative studies. Version September 9, 2025 submitted to Journal Not Specified21 of 25 The demonstrated effectiveness of ensemble approaches for aquatic species detection has broader implications for biodiversity monitoring and conservation efforts. The enhanced detection reliability and reduced error rates could significantly improve the accuracy of population assessments and ecological studies, leading to better-informed511 conservation decisions. The quantified trade-offs between accuracy and efficiency provide512 practical guidance for implementing these systems in field monitoring scenarios.

513 5.6. Limitations and Challenges514 Despite the promising results, several limitations must be acknowledged. First, the515 dataset, while comprehensive for Medaka species, is limited in scope compared to broader516 aquatic biodiversity. The generalizability of findings to other fish species or aquatic organ-517 isms requires further investigation. Second, the controlled and semi-controlled imaging518 conditions in our dataset may not fully represent the challenges encountered in completely519 natural field conditions.520 The computational overhead of ensemble methods represents a practical limitation521 for resource-constrained deployment scenarios. While this study has quantified these522 trade-offs, future work should explore optimization strategies to reduce computational re-523 quirements while maintaining accuracy benefits. Additionally, the storage and maintenance524 requirements for ensemble systems may present logistical challenges in field deployment525 scenarios.526 The temporal stability of ensemble performance across varying environmental con-527 ditions throughout different seasons and ecological cycles has not been fully evaluated.528 Long-term deployment studies would provide valuable insights into the robustness and529 maintenance requirements of ensemble-based monitoring systems.

530 5.7. Future Research Directions531 Several promising research directions emerge from this work. First, the exploration532 of lightweight ensemble architectures specifically designed for real-time ecological moni-533 toring applications could address current computational limitations. This could include534 investigation of knowledge distillation techniques to compress ensemble knowledge into535 more efficient single models.536 Second, the extension of ensemble approaches to multi-species detection and classifi- 537 cation tasks would provide broader applicability for biodiversity monitoring. This would538 require addressing challenges related to class imbalance, morphological similarity, and539 varying detection difficulty across species.540 Third, the integration of temporal information from video sequences could enhance541 detection reliability and enable behavior analysis capabilities. Ensemble approaches could542 be particularly effective for temporal fusion, combining spatial ensemble benefits with543 temporal consistency constraints.544 Finally, the development of adaptive ensemble systems that can dynamically adjust545 fusion strategies based on environmental conditions or image characteristics could optimize546 the accuracy-efficiency trade-off for specific deployment scenarios. This could include547 context-aware ensemble activation or confidence-based selective processing strategies.

548 6. Conclusion549 This research presents a comprehensive investigation of ensemble learning approaches550 for automated Medaka fish detection, demonstrating significant advances in both detection551 accuracy and methodological rigor for ecological monitoring applications. Through sys-552 tematic evaluation of YOLOv8-based ensemble frameworks employing Non-Maximum553 Suppression (NMS) and Weighted Boxes Fusion (WBF), we have established clear per-554 Version September 9, 2025 submitted to Journal Not Specified22 of 25 formance benchmarks and practical deployment guidelines for aquatic species detection555 systems.556 6.1. Key Findings and Contributions557 Our experimental results provide strong evidence for the superiority of ensemble558 approaches, with the WBF ensemble achieving a remarkable 21.1% improvement in559 mAP@0.5:0.95 compared to single-model baselines (0.5571 vs 0.4600). This improvement560 represents a substantial advancement in detection capability that directly translates to561 enhanced reliability for biodiversity monitoring applications. The 23.6% improvement562 in F1-score demonstrates the ensemble's superior balance between precision and recall,563 crucial for

minimizing both false positives and missed detections in ecological surveys.

The comprehensive comparison between NMS and WBF fusion strategies reveals that WBF's intelligent merging approach consistently outperforms traditional suppression methods, particularly in scenarios involving overlapping objects or uncertain boundaries common in aquatic environments. The 28.5% improvement in precision achieved by WBF over NMS ensemble highlights the value of advanced fusion strategies for ensemble object detection. The rigorous statistical validation, including cross-validation stability analysis and effect size quantification, establishes the reliability and generalizability of these performance improvements. The large effect sizes (Cohen's d > 1.0) observed for ensemble comparisons provide strong evidence for practical significance beyond statistical significance.

6.2. Practical Implications for Ecological Monitoring

The demonstrated accuracy improvements have direct implications for biodiversity monitoring and conservation efforts. The reduced error rates (from 15.3% to 8.9% for challenging scenarios) could significantly enhance the reliability of automated population assessments, leading to more accurate ecological insights and better-informed conservation decisions. The species-specific analysis reveals particular benefits for challenging species like O. javanicus, where precision improvements exceed 30% in optimal configurations.

The computational efficiency analysis provides essential guidance for practical deployment scenarios. While ensemble methods require approximately 4.3× more processing time, this trade-off is acceptable for many ecological monitoring workflows where accuracy is prioritized over real-time performance. The quantified throughput metrics (3,780 images/hour for WBF ensemble) indicate feasibility for large-scale archival image processing common in biodiversity surveys.

6.3. Methodological Advances

This work contributes several methodological advances to the field of automated ecological monitoring. The comprehensive evaluation framework, incorporating COCO-style metrics, cross-validation stability analysis, and statistical significance testing, establishes a rigorous standard for future comparative studies in this domain. The systematic confidence threshold analysis provides practical guidance for optimizing detection systems across different operational requirements. The detailed error analysis and failure case characterization offer valuable insights for understanding the limitations and optimal applications of different detection approaches. These findings inform both current deployment decisions and future research directions for improving automated species detection systems.

6.4. Limitations and Future Perspectives

While demonstrating significant advances, this research also reveals important limitations that warrant future investigation. The computational overhead of ensemble methods necessitates continued research into optimization strategies, including lightweight ensemble architectures and selective activation mechanisms. The dataset scope, while comprehensive for Medaka species, requires extension to broader aquatic biodiversity for generalized conclusions. Future research directions include the development of real-time ensemble systems through architectural optimization, extension to multi-species detection scenarios, integration of temporal information from video sequences, and exploration of adaptive ensemble systems that dynamically optimize performance based on environmental conditions.

6.5. Broader Impact and Significance

The successful application of advanced ensemble learning techniques to ecological monitoring represents a significant step toward more reliable automated biodiversity assessment systems.

The demonstrated improvements in detection accuracy and reliability could accelerate the adoption of computer vision technologies in conservation efforts, enabling larger-scale and more cost-effective monitoring programs.

The rigorous methodological framework established in this work provides a foundation for future research in automated ecological monitoring, while the practical deployment insights facilitate real-world implementation of these technologies. As computational resources continue to advance and optimization techniques improve, the accuracy benefits demonstrated here will become increasingly accessible for field deployment scenarios.

In conclusion, this research establishes ensemble learning as a valuable approach for enhancing automated aquatic species detection, providing both immediate practical benefits and a foundation for continued advancement in this critical application domain. The demonstrated improvements in detection reliability, combined with comprehensive methodological validation, represent a significant contribution to the intersection of computer vision and ecological science, with direct implications for biodiversity conservation and ecosystem monitoring efforts.

Funding: This research was supported by the Data Science and Artificial Intelligence Research Group at Hasanuddin University and the B.J. Habibie Institute of Technology. Computational resources were provided by the Faculty of Mathematics and Natural Sciences, Hasanuddin University. No external commercial funding was received for this research.

Acknowledgments: The authors express their sincere gratitude to the marine biology research teams at Hasanuddin University for their expertise in fish species identification and annotation protocols. We thank the Data Science and Artificial Intelligence Research Group for providing computational resources and technical support throughout this research. Special appreciation is extended to the field researchers who contributed to the diverse image collection efforts across multiple aquatic environments in Sulawesi and Java. The authors acknowledge the valuable collaboration with the Kyushu Institute of Technology for methodological guidance and validation protocols. We also thank the open-source community for developing and maintaining the software tools that made this research possible, including the YOLOv8 framework, PyTorch, and associated computer vision libraries. The ichthyological expertise provided by the Department of Biology and Department of Fishery at Hasanuddin University was instrumental in ensuring accurate species identification and annotation quality.

Conflicts of Interest: The authors declare no conflicts of interest. The research was conducted independently without commercial partnerships or competing interests that could influence the interpretation of results.

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521, 436–444. https://doi.org/10.1038/nature14539.

2.Zhao, Z.Q.; Zheng, P.; Xu, S.t.; Wu, X. Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems 2019, 30, 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865.
3. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587. https://doi.org/10.1109/CVPR.2014.81.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems 2015, 28, 91–99. https://doi.org/10.1109/TPAMI.2016.2577031.
5. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

https://doi.org/10.1109/CVPR.2016.91.

6. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7263–7271. https://doi.org/10.1109/CVPR.2017.690.

7. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934 2020.

8. Terven, J.; Cordova-Esparza, D. A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and beyond. Machine Learning and Knowledge Extraction 2023, 5, 1680–1716. https://doi.org/10.3390/make5040083.

9. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; et al. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation. Zenodo 2022. https://doi.org/10.5281/zenodo.3908559.

10.Dietterich, T.G. Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems 2000, pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1.

11.Zhou, Z.H.; Wu, J.; Tang, W. Ensembling neural networks: many could be better than all. Artificial Intelligence 2002, 137, 239–263. https://doi.org/10.1016/S0004-3702(02)00190-X.

12.Breiman, L. Random forests. Machine Learning 2001, 45, 5–32. https://doi.org/10.1023/A:1010933404324.

13.Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. In Proceedings of the Journal of Computer and System Sciences. Elsevier, 1997, Vol. 55, pp. 119–139. https://doi.org/10.1006/jcss.1997.1504.

14.Solovyev, R.; Wang, W.; Gabruseva, T. Weighted Boxes Fusion: Ensembling Boxes from Different Object Detection Models. Image and Vision Computing 2021, 117, 104–127. https://doi.org/10.1016/j.imavis.2021.104127.

15. Kalafi, E.; Javanmard, M. A Review on Deep Learning Approaches in Underwater Image Processing. International Journal of Computer Vision and Image Processing 2018, 8, 1–15. https://doi.org/10.4018/IJCVIP.2018010101.

16.Leow, W.K.; Savariar, B. Challenges and Techniques in Underwater Imaging. In Proceedings of the 2015 International Conference on Underwater Systems Technology: Theory and Applications (USYS), 2015, pp. 1–5. https://doi.org/10.1109/USYS.2015.7440944.

17.Qin, H.; Li, X.; Liang, J.; Peng, Y.; Zhang, C. DeepFish: Accurate underwater live fish recognition with a deep architecture. Neurocomputing 2016, 187, 49–58. https://doi.org/10.1016/j.neucom.2015.10.122.

18.Mandal, R.; Connolly, R.M.; Schlacher, T.A.; Stantic, B. Assessing fish abundance from underwater video using deep neural networks. ICES Journal of Marine Science 2018, 75, 1526–1535. https://doi.org/10.1093/icesjms/fsy038.

19.Ditria, E.M.; Connolly, R.M.; Jinks, K.J.; Lopez-Marcano, S. Annotated video footage for automated identification and counting of fish in unconstrained environments. Scientific Data 2020, 7, 1–7. https://doi.org/10.1038/s41597-020-0465-9.

20. Campbell, M.D.; Pollack, A.G.; Gledhill, C.T.; Switzer, T.S.; DeVries, D.A. Comparison of relative abundance indices calculated from two methods of generating video count data. Fisheries Research 2015, 170, 125–133. https://doi.org/10.1016/j.fishres.2015.05.011.

21.Tamou, A.B.; Benzinou, A.; Nasreddine, K.; Ballihi, L. Underwater live fish recognition by deep learning. Ecological Informatics 2021, 63, 101322.

https://doi.org/10.1016/j.ecoinf.2021.101322.

22. Salimi, M.; Bai, Y. Real-time fish detection and tracking in underwater videos based on deep learning. Neurocomputing 2016, 275, 1–12. https://doi.org/10.1016/j.neucom.2017.10.010.

23. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162. https://doi.org/10.1109/CVPR.2018.00644.

24. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. Computer Vision and Pattern Recognition (CVPR) 2016, pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.

25. Stone, M. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological) 1974, 36, 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

26. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection.Proceedings of the 14th International Joint Conference on Artificial Intelligence 1995, 2, 1137–1143.

27. Browne, M.W. Cross-validation methods. Journal of Mathematical Psychology 2000, 44, 108–132. https://doi.org/10.1006/jmps.1999.1279.

28. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. Journal of Big Data 2019, 6, 1–48. https://doi.org/10.1186/s40537-019-0197-0.