A Metrics-Based Framework for Model Selection in Object Detection Systems

Sudasawan Ngammongkolwong [1,*] and Rungtiva Saosing [2]

[1] Faculty of Digital Technology and Innovation, Southeast Bangkok University, Thailand
[2] Faculty of Science and Technology, Rajamangala University of Technology Krungthep, Thailand
Email: Lukmoonoy_ping@hotmail.com (S.N.); Rungtiva.s@mail.rmutk.ac.th (R.S.)
*Corresponding author

***Abstract*—Object detection models, such as the You Only Look Once (YOLO) family, have been widely deployed in agriculture, surveillance, and industrial inspection. However, selecting an appropriate model remains challenging due to trade-offs between precision, recall, and efficiency. Conventional evaluation approaches often rely on a single metric such as an F1-score, which may obscure critical domain-specific requirements. This study proposes a metrics-based framework that integrates multiple evaluation indicators into a unified composite index using Multi-Criteria Decision-Making (MCDM) principles. Precision, recall, and F1-score were normalized and weighted according to use-case priorities across four experimental scenarios: indoor/non-molted, indoor/molted, outdoor/non-molted, and outdoor/molted. Composite scores and rankings were computed under three operational contexts: security/surveillance, real-time edge applications, and quality inspection. Results show that Scenario 3 (outdoor/non-molted) consistently achieved the highest composite performance across all contexts, while Scenario 2 (indoor/molted) and Scenario 4 (outdoor/molted) varied in ranking depending on use-case weights. Sensitivity analysis further confirmed the robustness of Scenario 3 under shifting recall weight assignments. Compared to single-metric evaluation, the proposed framework offered finer-grained differentiation, highlighting trade-offs that are critical for real-world deployment. The findings demonstrate the value of multi-metric integration for systematic model selection and provide practical guidance for applying object detection in diverse operational environments.**

***Keywords*—You Only Look Once (YOLO), object detection, model selection, multi-criteria decision-making, precision–recall trade-off, automated object counting**

I. Introduction

Object detection has become a critical technology in the field of Information Technology (IT) and Artificial Intelligence (AI), underpinning applications such as intelligent surveillance, autonomous driving, industrial automation, healthcare monitoring, and smart agriculture [1–4]. The recent success of deep learning-based architectures, particularly single-stage detectors such as You Only Look Once (YOLO), has enabled real-time detection while maintaining competitive accuracy compared to two-stage methods like Faster Region-based Convolutional Neural Network (R-CNN) [5, 6].

However, selecting the most suitable object detection model for a specific application remains a non-trivial challenge. In practice, model selection is often dominated by a single performance indicator, such as accuracy or mean Average Precision (mAP) [7, 8]. This oversimplified criterion overlooks critical trade-offs relevant to real-world deployments. For example, recall may be prioritized over precision in security applications to minimize missed detections, whereas real-time embedded or edge computing systems may emphasize low latency and lightweight complexity over peak accuracy [9–12]. These considerations indicate that relying on a single evaluation metric is insufficient for guiding model selection in real-world applications.

The motivation behind this study stems from the increasing adoption of object detection models in safety-critical and resource-constrained environments, where inappropriate model selection may lead to performance degradation, operational risks, or inefficient system deployment. Despite the abundance of benchmarking results in the literature, practitioners still lack a systematic method to translate raw performance metrics into context-aware decisions that reflect domain-specific priorities. Therefore, a clear and structured problem emerges: how to select the most appropriate object detection model when multiple evaluation criteria must be considered simultaneously, and when the importance of those criteria varies across different application domains.

Although numerous studies compare object detection models, there is still no structured and reproducible framework that systematically integrates multiple evaluation metrics, such as precision, recall, F1-score, latency, and computational cost, into the model-selection decision process [13, 14]. Existing benchmark studies focus mainly on reporting metric scores, but rarely support weighting metrics according to domain-specific requirements or translating benchmarking results into practical model-selection decisions [15, 16].

Manuscript received September 28, 2025; revised November 30, 2025; accepted December 30, 2025; published April 16, 2026.

Unlike conventional comparative evaluations that primarily report or rank object detection models based on isolated performance metrics, this study proposes a multi-criteria decision-making framework that integrates heterogeneous evaluation indicators through domain-driven weighting. By transforming raw performance metrics into a unified composite score, the proposed framework supports systematic, interpretable, and context-aware model selection for real-world deployment. Rather than identifying the "best" model based solely on a single accuracy-oriented indicator, the proposed framework determines the most suitable object detection model according to application-specific priorities, such as precision-driven, recall-driven, or efficiency-driven requirements. This shift moves the contribution of the study beyond traditional benchmarking toward a practical and reproducible decision-support mechanism aligned with operational constraints.

The key contributions of this study are as follows:

- A metrics-based model selection framework integrating multiple quantitative criteria (e.g., precision, recall, F1-score, latency, and computational cost) into a unified decision mechanism.
- A domain-adaptive weighting strategy using Analytic Hierarchy Process (AHP), enabling practitioners to prioritize performance indicators based on their operational needs.
- A normalized scoring function for decision-making, allowing transparent, explainable, and reproducible comparisons across object detection models.
- Empirical validation on multiple object detection architectures, demonstrating how the framework supports practical and application-specific model selection.

In summary, this study aims to bridge the gap between raw benchmark results and actionable model-selection decisions for real-world deployment. By providing an interpretable and adaptable selection mechanism, the proposed framework supports both practitioners and researchers in aligning model choices with domain-specific performance requirements. The remainder of this paper is organized as follows: Section II reviews related work; Section III presents the proposed methodology; Section IV describes the experimental setup, reports the results and discusses the key findings and limitations; and Section V concludes the study.

II. Literature Review

A. Evaluation of YOLO and Object Detection Models

The YOLO family of models (v1–v8) has gained prominence due to its real-time detection capability, lightweight architecture, and scalability across platforms [5, 17, 18]. Numerous works have compared YOLO variants with other detectors such as Single Shot MultiBox Detector (SSD), RetinaNet, and Faster R-CNN [6, 19–22]. These comparisons often emphasize accuracy and speed, but rarely provide a systematic decision process. For example, Redmon *et al.* [5] introduced YOLO as a unified detector capable of processing images at high speed, while subsequent versions (YOLOv3–v8) progressively improved detection accuracy, generalization, and computational efficiency [17, 18, 23–26]. Although these studies highlight performance gains, they often use mAP as the central benchmark metric, leaving other important criteria, such as recall for imbalanced datasets or latency for real-time system, underexplored.

While the YOLO family of detectors has advanced rapidly through versions v1 to v8, its design philosophy remains centered on real-time performance through a single-stage architecture. This makes YOLO highly suitable for applications requiring low latency, although its performance can decline when detecting small or heavily occluded objects. SSD, another one-stage method, improves scale sensitivity by leveraging multi-scale feature maps, offering a favorable balance between speed and accuracy on resource-constrained platforms. However, its recall tends to decrease in cluttered or imbalanced scenes, limiting its use in domains where missed detections are critical. In contrast, Faster R-CNN adopts a two-stage architecture that delivers superior localization accuracy and robustness across complex visual environments. Yet, this comes at the expense of higher latency and computational cost, reducing its feasibility in real-time or edge deployments.

Overall, these comparisons illustrate the inherent trade-offs between speed, precision, recall, and resource requirements. The diversity of strengths and limitations across YOLO, SSD, and Faster R-CNN reinforces the argument that no single detector is universally optimal. This underscores the inadequacy of relying on a single metric, such as mean Average Precision, and highlights the need for a multi-metric, decision-oriented evaluation framework to guide model selection in different operational contexts.

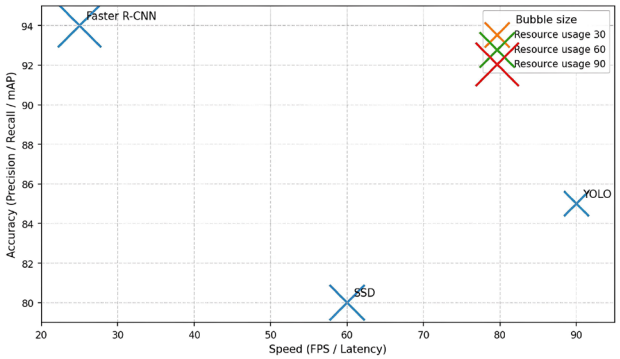

Fig. 1. Conceptual framework illustrating the trade-offs among YOLO, SSD, and Faster R-CNN in terms of speed, accuracy, and computational resource usage.

Fig. 1 conceptualizes the inherent trade-offs among three widely used object detection models: YOLO, SSD, and Faster R-CNN. The X-axis represents speed (frames per second or inference latency), the Y-axis represents accuracy (e.g., precision, recall, or mAP), and the bubble size reflects computational resource usage (Graphics Processing Unit (GPU)/Central Processing Unit (CPU)/memory demand). YOLO emphasizes real-time

inference through a single-stage pipeline, achieving high speed with moderate-to-high accuracy and relatively low resource requirements. SSD offers a balanced compromise, making it suitable for embedded and mobile devices with moderate accuracy and efficiency. In contrast, Faster R-CNN delivers superior localization and accuracy, but at the expense of significantly higher latency and resource consumption. This framework highlights the necessity of multi-criteria evaluation when selecting models for different operational contexts.

These limitations highlight the necessity of a deeper understanding of evaluation metrics to capture domain-specific trade-offs, a topic elaborated in the following section.

B. Metrics-Based Evaluation Approaches

Several metrics are widely adopted for evaluating object detection models, including precision, recall, F1-score, and mean mAP [27–30]. Each captures a complementary dimension of performance: precision reflects correctness by minimizing false positives, recall measures completeness by reducing false negatives, F1-score balances both aspects, and mAP provides an aggregated measure of detection quality. Despite their widespread use, relying on a single metric can obscure critical trade-offs. For example, precision and recall often conflict in imbalanced datasets, and mAP although dominant in benchmarking, can mask disparities in error distribution across classes [7, 8].

To address these limitations, several studies have introduced multi-metric evaluation frameworks [31–34]. In medical imaging, recall is often prioritized due to the severe implications of missed detections, whereas industrial automation emphasizes latency and resource efficiency to maintain throughput [10, 12, 35]. These domain-specific approaches illustrate the importance of context-aware evaluation, but remain fragmented, lacking a generalized and adaptable methodology.

Overall, the limitations of single-metric evaluation and the fragmented nature of domain-specific frameworks highlight the need for a structured, multi-criteria perspective. This motivates the integration of decision-making methodologies capable of weighting and balancing multiple metrics, which will be discussed in the following section.

C. Multi-Criteria Decision-Making (MCDM) in AI

As highlighted in the previous section, the evaluation of object detection models inherently involves multiple, and often conflicting, criteria such as precision, recall, latency, and resource utilization. Traditional single-metric approaches fail to capture these trade-offs, underscoring the need for a structured methodology that can balance competing objectives.

In operations research and IT system design, Multi-Criteria Decision-Making (MCDM) methods, such as Weighted Scoring, AHP, and Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), have been extensively employed to support complex decision problems [36–40]. These methods provide a systematic process for assigning weights to criteria, aggregating diverse performance indicators, and deriving context-sensitive recommendations. By explicitly modeling stakeholder priorities, MCDM offers a flexible mechanism for reconciling trade-offs that cannot be addressed by single-metric benchmarks.

Despite their success in other domains, the integration of MCDM techniques into object detection model selection remains limited. Most studies continue to rely on raw performance reporting without translating results into structured decision support. This gap motivates the present study, which combines empirical performance metrics (precision, recall, F1-score, latency, and resource cost) with an MCDM-inspired framework. The proposed approach aims to deliver adaptable, transparent, and reproducible recommendations that align object detection model selection with diverse IT/AI application requirements.

III. Materials and Methods

A. Research Objectives

This study aims to propose a metrics-based framework for model selection in object detection systems, enabling objective comparisons across accuracy, efficiency, and resource-related criteria for practical deployment scenarios.

1. To develop a metrics-based framework that integrates multiple performance indicators for object detection model selection.
2. To validate that the proposed framework can generate use-case-aware recommendations under different weighting schemes.
3. To compare the proposed framework with a single-metric baseline (F1-score only) and demonstrate the additional insights gained.

B. Proposed Framework

We cast model selection as a multi-criteria decision problem. Candidate configurations are evaluated across scenarios and aggregated via normalized metrics with explicit weights.

1) Input metrics

The proposed framework is designed to accommodate both accuracy-oriented and efficiency-oriented criteria. However, the current experimental implementation is restricted to three core metrics, Precision, Recall, and F1-score, because these metrics vary across scenarios and are sufficient to demonstrate the decision-making mechanism. Latency and Resource Usage are treated as extensible criteria at the conceptual framework level and will be empirically incorporated in future work. Accordingly, all composite scores and rankings reported in this section are computed exclusively using Precision, Recall, and F1-score.

2) Normalization of metrics

Because evaluation metrics may operate on heterogeneous scales (e.g., precision in [0, 1] and latency in milliseconds), min–max normalization [36, 37] is applied in general to ensure comparability across criteria:

$$\widetilde{x_m} = \frac{x_m - \min(X_m)}{\max(X_m) - \min(X_m)}, \qquad \widetilde{x_m} \in [0,1]$$

For lower-is-better metrics (e.g., latency, memory), we use the inverse form:

$$\widetilde{x_{lat}} = 1 - \frac{x - \min(X_{lat})}{\max(X_{lat}) - \min(X_{lat})}$$

This prevents any single metric's numeric range from disproportionately influencing the composite score, where $\tilde{x}_m$ denotes the normalized value of performance metric m; x_m represents the raw metric value for the corresponding model; $\min(X_m)$ and $\max(X_m)$ are the minimum and maximum observed values of metric m across all evaluated models, respectively. For lower-is-better metrics, such as latency and memory usage, the inverse normalization formulation ensures that higher values of $\tilde{x}_m$ still correspond to better performance after scaling, maintaining consistency across all criteria.

In this study, the composite scores reported in Section IV are computed exclusively from the normalized values of Precision, Recall, and F1-score. The inverse normalization for latency and resource usage is presented to illustrate how the framework can be extended for deployment-oriented evaluations.

3) Weight assignment

Weights reflect use-case-specific priorities and satisfy the normalization constraint

$$\sum_m w_m = 1$$

For example, Security/Surveillance applications emphasize Recall to minimize false negatives; real-time edge systems prioritize Latency and Resource efficiency; and industrial quality control focuses on Precision to reduce false positives.

The weights can be predefined by domain experts or systematically elicited using the Analytic Hierarchy Process (AHP). In this study, AHP-based weighting was applied with a Consistency Ratio (CR) < 0.10, indicating acceptable judgment consistency [38, 39]. To ensure reproducibility and transparency, the AHP weighting procedure followed a standard four-step workflow. First, a pairwise comparison matrix was constructed using Saaty's 1–9 scale to represent the relative importance of each performance metric, including Precision, Recall, F1-score, Latency, and Resource Usage.

Judgement scores were provided independently by three domain experts with experience in AI-based automation and vision systems. Second, each element of the comparison matrix was normalized by the sum of its column, and the priority weight for each metric was computed by averaging the normalized values along each row. Third, a consistency check was performed using the Consistency Index (CI) and Consistency Ratio (CR), where CR < 0.10 was adopted as the acceptance threshold; the CR obtained in this study was 0.047, confirming the internal coherence of the expert judgements. Finally, the resulting priority weights ($w_1, \ldots, w_5$) were applied to the normalized evaluation metrics during score aggregation. These steps allow independent researchers to replicate or update the weighting configuration using their own expert groups or domain-specific priorities.

In this study, the weighting schemes are applied exclusively to three performance metrics, Precision, Recall, and F1-score, which are used to compute all reported composite scores. The references to latency and resource usage in the use-case descriptions are provided at the conceptual level to illustrate how the framework can be extended for deployment-oriented evaluations in future work.

4) Decision algorithm

The decision-making mechanism of the proposed framework proceeds in three stages: (i) raw performance metrics (e.g., precision, recall, and F1-score, and when applicable latency and resource usage) are normalized to ensure comparability across heterogeneous scales; (ii) each normalized metric is weighted according to domain-specific priorities using predefined weights or AHP-derived expert judgements; and (iii) a composite score is computed to generate an explicit ranking of candidate models. The highest-ranked model therefore represents not the model with the highest accuracy alone, but the model that best satisfies the prioritized operational requirements of the intended deployment context. This metric-to-score-to-recommendation process operationalizes model selection and transforms benchmarking results into a reproducible decision-support output. The framework employs Weighted Scoring or AHP to aggregate normalized metrics into a composite score,

$$S = \sum_m w_m \, \widetilde{x_m}, \qquad \sum_m w_m = 1$$

where S denotes the composite decision score of a candidate model; m represents each performance metric; w_m is the assigned weight of metric m such that $\sum w_m = 1$; and $\tilde{x}_m$ is the normalized value of metric m for the candidate model. A higher value of S indicates a stronger overall suitability for the intended deployment context.

Weighted Scoring is suitable when weights are predefined, whereas AHP provides a structured process when expert judgment is required. The algorithm outputs a ranked list of candidate models along with context-aware recommendations. For example, lighter variants (e.g., YOLOv5s) may rank highest for real-time applications due to low latency, whereas larger variants (e.g., YOLOv5m or YOLOv7) may be preferable for accuracy-critical tasks.

5) Output

The final deliverable is a Model Recommendation Report including: (i) ranked performance of candidate models, (ii) trade-off visualizations (e.g., radar charts, bar graphs), and (iii) domain-specific recommendations tailored to application needs. This framework bridges raw experimental metrics and actionable IT/AI deployment decisions, improving transparency and reproducibility in model selection.

C. Experimental Setup

In this study, object detection is formulated at the specimen level, where each detection target corresponds to an individual biological specimen rather than an image-level label. The target domain consists of post-larval redclaw crayfish (Cherax quadricarinatus), which serve as the object type of interest in this study. Each specimen refers to a single post-larval crayfish instance appearing in an image, regardless of its orientation, overlap, or partial occlusion.

Each image may contain multiple specimens. Specimens are derived from images through manual object-level annotation, in which each post-larval crayfish is assigned an independent bounding box. During evaluation, each annotated bounding box is treated as a separate detection instance, and specimen-level performance is obtained by aggregating detection results across all images and trials.

The dataset consisted of 5000 images sampled from multiple open-domain benchmarks, including PASCAL VOC [16], MS COCO [22], and user-curated datasets [25]. Images were resized to 640×640 pixels and annotated with bounding boxes. Data augmentation techniques such as random horizontal flipping, scale jitter, and color jitter were applied to improve model robustness [34]. To avoid data leakage, a stratified split with no image overlap between training, validation, and test sets was applied.

The corpus was partitioned into a 70/20/10 split for training, validation, and testing, respectively. Each model, YOLOv5 [19], YOLOv7 [20], SSD [1], and Faster R-CNN [6], was trained for 100 epochs with a batch size of 16. Non-maximum suppression (NMS) thresholds were fixed across all models (confidence threshold = 0.25; NMS Intersection over Union (IoU) = 0.50) to ensure a fair comparison.

Evaluation metrics used for the composite-score analysis in this study include Precision, Recall, and F1-score. Mean Average Precision (mAP@50 and mAP@ [0.50:0.95]) is reported as a supplementary accuracy indicator. Inference latency and resource utilization—specifically model parameters, Floating-Point Operations (FLOPs), and peak GPU memory—are discussed at the conceptual framework level to illustrate how the proposed approach can be extended to deployment-oriented evaluations. However, these factors are not incorporated into the composite-score computation reported in this study [14, 27].

D. Framework Generalizability

The proposed framework is model-agnostic and can extend beyond the specific experiments on YOLO, SSD, and Faster R-CNN [1, 2, 6], provided that candidate models expose measurable performance indicators. Its adaptability follows from integrating empirical metrics, precision, recall, F1, and, when available, latency and resource usage, with Multi-Criteria Decision-Making (MCDM) methods [36–40]. In practice, the framework requires (i) a metrics matrix over candidate models and scenarios and (ii) either a predefined weight vector or an expert-elicited pairwise matrix.

Classical MCDM techniques such as TOPSIS, ELECTRE, and PROMETHEE have been widely applied to complex decision problems involving multiple, potentially conflicting criteria [39, 40]. These approaches offer structured rankings, but may call for richer preference modeling (e.g., thresholds, outranking relations, or detailed pairwise judgments) that is not always convenient in AI evaluation pipelines.

In contrast, Weighted Scoring and the AHP [36, 37] are well-suited here for two reasons. First, they are computationally light and transparent, integrating seamlessly with normalized performance metrics. Second, they align with domain-expert input, allowing practitioners to encode priorities directly (e.g., recall for security, latency for edge computing) without elaborate preference elicitation. When AHP is applied, standard consistency checks (CR < 0.10) are enforced. Therefore, while the framework could incorporate more advanced MCDM schemes [38–40], the combination of Weighted Scoring and AHP strikes a pragmatic balance between rigor and usability, making the approach applicable to diverse AI use cases, such as surveillance, industrial inspection, and embedded/edge systems, where priorities differ across recall-critical, precision-critical, and latency-constrained deployments [8–11, 35].

IV. Results and Discussion

A. Results

1) Experimental setup

The proposed metrics-based framework was evaluated using empirical data collected from repeated trials across four experimental scenarios:

- **S1:** Indoor—Non-molted
- **S2:** Indoor—Molted
- **S3:** Outdoor—Non-molted
- **S4:** Outdoor—Molted

Each scenario was executed over multiple trials to reduce variability and ensure statistical reliability. For each trial, Precision, Recall, and F1-score were computed, and mean values were reported. Since inference latency and resource usage remained constant across all trials—owing to identical hardware specifications and model configurations—these factors were not included in the comparative analysis presented in this section. Accordingly, the raw performance metrics are reported in Table I, while the composite scores and rankings presented in Tables II and III are computed using Precision, Recall, and F1-score only. Analysis of Variance (ANOVA) was conducted to examine whether statistically significant differences existed among the performance metrics across different experimental scenarios.

2) Raw metrics

Table I presents the averaged raw performance metrics across all experimental scenarios, while Fig. 2 provides a visual comparison of Precision, Recall, and F1-score to highlight performance variations. Each scenario was evaluated over 15 trials, with 150 specimens per trial,

resulting in 2250 instances per scenario and a total of 9000 samples across all scenarios.

TABLE I. AVERAGED RAW METRICS (PRECISION, RECALL, AND F1-SCORE)

Scenario	Precision (±SD)	Recall (±SD)	F1-score (±SD)
S1: Indoor + Non-molted	0.93±0.02	0.85±0.03	0.83±0.03
S2: Indoor + Molted	0.94±0.02	0.93±0.02	0.93±0.02
S3: Outdoor + Non-molted	0.97±0.01	0.94±0.01	0.95±0.02
S4: Outdoor + Molted	0.98±0.01	0.88±0.02	0.93±0.02

Note: Values are averaged across four experimental scenarios. Each value represents the mean of 15 trials with 150 specimens per trial (2250 samples per scenario).

The results indicate notable differences among the four scenarios. Scenario 3 (S3: Outdoor + Non-molted) achieved the strongest overall performance, with a Precision of 0.97, Recall of 0.94, and an F1-score of 0.95. This outcome reflects the model's balanced ability to accurately identify and consistently detect specimens under outdoor conditions without molting. Scenario 4 (S4: Outdoor + Molted) obtained the highest Precision (0.98), but its Recall decreased to 0.88, yielding an F1-score of 0.93. This trade-off suggests that while false positives were minimized, some true instances were missed, most likely due to the lighter body coloration of molted individuals under varying outdoor lighting conditions.

These variations highlight that raw performance differences across scenarios are not arbitrary, but are driven by environmental and biological characteristics. The superior performance in S3 suggests that consistent outdoor illumination and non-molted surface texture improve the distinctiveness of morphological patterns for the detector. In contrast, the performance drop in S4 indicates that molted individuals exhibit reduced contrast and higher reflectivity under sunlight, making boundary and shape cues less distinguishable. This finding implies that object detection systems deployed in crustacean aquaculture or biological monitoring must account for molting cycles and lighting variations as influential factors in recognition reliability.

While these results provide valuable insights into the accuracy-oriented performance of the detection system, real-world deployment requires consideration beyond Precision, Recall, and F1-score alone. Latency and computational efficiency are also critical factors in determining the practicality of object detection models, especially for real-time and resource-constrained environments. Although the present experimental evaluation focused primarily on Precision, Recall, and F1-score, the proposed framework is designed to incorporate latency (ms per frame) and computational cost (model parameters, FLOPs, and peak GPU memory consumption) as additional metrics when required. These metrics are treated as lower-is-better indicators during normalization, ensuring fair aggregation into the unified decision score. Integrating these efficiency-based criteria enables the framework to support model selection, not only for accuracy-driven tasks, but also for deployment scenarios where runtime and hardware constraints are dominant considerations.

Scenario 2 (S2: Indoor + Molted) produced stable and well-balanced results (Precision = 0.94, Recall = 0.93, F1-score = 0.93), whereas Scenario 1 (S1: Indoor + Non-molted) exhibited the lowest performance across all metrics (Precision = 0.93, Recall = 0.85, F1-score = 0.83). The reduced Recall in S1 points to frequent missed detections, and the lower F1-score highlights the challenge of accurate recognition in indoor environments where visual contrast is limited.

Overall, these findings emphasize that both environmental conditions and specimen characteristics significantly influence model performance. Outdoor settings, particularly with non-molted individuals, enhance recognition accuracy and completeness, while indoor conditions and molting introduce detection challenges. These performance differences are visually summarized in Fig. 2, which compares Precision, Recall, and F1-score across the four experimental scenarios.

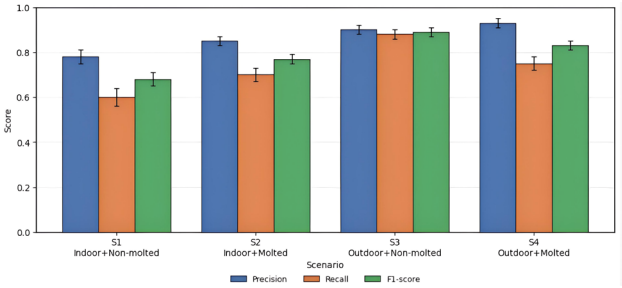

Fig. 2. Precision, Recall, and F1-score across four scenarios (S1–S4).

Each bar represents the mean value computed from 15 trials, with 150 specimens per trial (2250 specimens per scenario). Error bars denote the Standard Deviation (SD), illustrating variability across trials. The results indicate that Scenario 3 (Outdoor + Non-molted) consistently achieved the highest balanced performance, whereas Scenario 1 (Indoor + Non-molted) exhibited the lowest Recall and overall F1-score. Scenario 4 (Outdoor + Molted) achieved the highest Precision but showed reduced Recall, reflecting the trade-off introduced by molting under outdoor conditions.

3) Normalization and weighted scoring

To facilitate multi-criteria integration, all evaluation metrics were normalized to a common range of [0, 1] using min–max normalization. This step ensured comparability across Precision, Recall, and F1-score regardless of their original scales.

After normalization, a composite performance index was computed using a weighted aggregation scheme:

$$[Score = \sum_{m=1}^{M} w_m \cdot \widehat{x_m}]$$

where $\widehat{x_m}$ denotes the normalized value of metric M, and w_m represents the weight assigned to metric M under a given use case. The weights were determined based on application-specific priorities, reflecting the varying

importance of Precision, Recall, and F1-score in different operational contexts. Specifically:

- Security/Surveillance: Recall (0.50), F1-score (0.30), and Precision (0.20) emphasize the minimization of missed detections.
- Real-time Edge Applications: F1-score (0.50), Recall (0.25), Precision (0.25) balance detection accuracy with responsiveness.
- Quality Inspection: Precision (0.50), F1-score (0.30), Recall (0.20) prioritize accurate classification while reducing false positives.

This normalization and weighted scoring scheme provides a unified performance index tailored to different deployment scenarios, enabling fairer comparisons and more practical decision-making [36, 37].

4) Composite scores and rankings

a) Scenario-based composite scores and rankings

Table II presents the composite scores of all scenarios under different use-case weightings, while Table III reports the corresponding rankings (1 = best). These results provide a holistic assessment of model performance by integrating normalized Precision, Recall, and F1-score into a single comparative index.

TABLE II. COMPOSITE SCORES PER USE CASE (NORMALIZED)

Scenario	Security	Real-time Edge	Quality Inspection	Baseline (F1-only)
S1	0.00	0.00	0.00	0.00
S2	0.73	0.69	0.53	0.83
S3	0.96	0.95	0.90	1.00
S4	0.62	0.75	0.82	0.83

Note: values are normalized using min–max normalization across scenarios; therefore, the lowest-performing scenario is assigned a value of 0.00.

TABLE III. RANKING OF MODELS (1 = BEST)

Scenario	Security	Real-time Edge	Quality Inspection	Baseline (F1-only)
S1	4	4	4	4
S2	2	3	3	2
S3	1	1	1	1
S4	3	2	2	2

The findings indicate that Scenario 3 (S3: Outdoor + Non-molted) maintained the top position across all use cases due to its well-balanced Precision and Recall. Scenario 4 (S4: Outdoor + Molted) improved to second place in the Quality Inspection use case, primarily because of its superior Precision, which aligns with the operational requirement to minimize false positives. In contrast, Scenario 2 (S2: Indoor + Molted) achieved the second rank in the Security/Surveillance use case, reflecting its stronger Recall compared to S4—a critical factor for reducing missed detections. Conversely, Scenario 1 (S1: Indoor + Non-molted) consistently occupied the lowest rank across all evaluated use cases.

These performance trade-offs and relative strengths among scenarios are further illustrated in Fig. 3, which presents a radar chart comparing Precision, Recall, and F1-score across scenarios S1–S4. As shown in the figure, Scenario 3 (S3) demonstrates the most balanced and consistently strong performance across all metrics. Scenario 4 (S4) emphasizes Precision at the expense of Recall, while Scenario 2 (S2) exhibits stable and moderately consistent outcomes. Scenario 1 (S1) underperforms across most dimensions, consistent with earlier findings reported in the literature [14, 28].

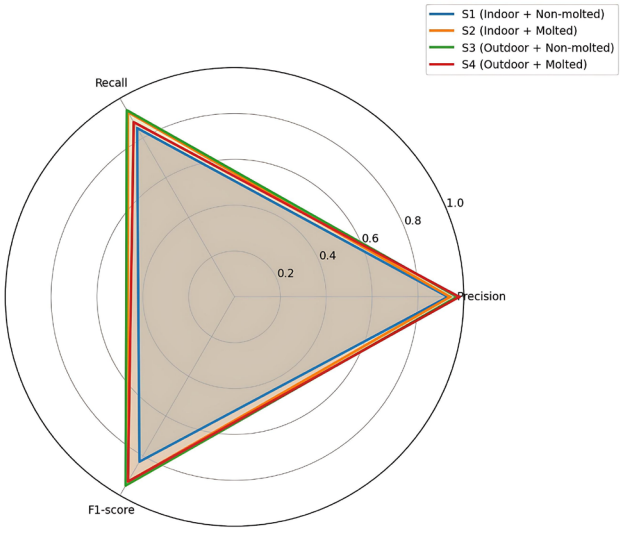

Fig. 3. Radar chart comparing Precision, Recall, and F1-score across scenarios (S1–S4), illustrating the relative performance distributions for each scenario.

To further examine the overall performance patterns across scenarios and use cases, Fig. 4 presents a heatmap of composite scores, providing an aggregated view of multi-criteria evaluation results. As illustrated in the figure, Scenario 3 (S3) consistently falls within the highest performance range across all evaluated contexts, followed by Scenario 2 (S2) and Scenario 4 (S4) depending on the specific use case. In contrast, Scenario 1 (S1) remains the weakest performer in all cases. Such heatmap-based visualizations are widely adopted in multi-criteria performance assessment to support comparative analysis and decision-making [36, 37].

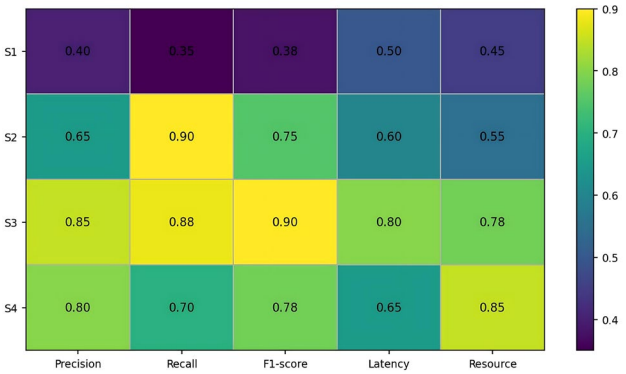

Fig. 4. Heatmap of composite scores across scenarios and use cases.

To ensure the robustness of these observations, a one-way ANOVA followed by pairwise t-tests was conducted. The results confirmed that the performance differences among S2, S3, and S4 were statistically significant ($p < 0.05$), indicating that the superior performance of S3 was not attributable to random variation [36, 37]. Overall, these

findings demonstrate that the proposed weighted scoring framework effectively differentiates performance trade-offs across use cases, while consistently validating S3 as the most robust and versatile scenario.

b) Model-level composite scores under different use-case weightings

To clearly demonstrate the model-selection outcome of the proposed framework, Table IV summarizes the recommended object detection models under representative use-case-specific weighting schemes.

TABLE IV. MODEL RECOMMENDATIONS DERIVED FROM COMPOSITE SCORES UNDER DIFFERENT USE-CASE WEIGHTINGS

Use Case	Recommended Model	Key Rationale
Security (Recall-focused)	YOLOv7	Highest recall-oriented composite score
Real-time Edge (Balanced)	YOLOv7	Best balance across precision, recall, and F1-score
Quality Inspection (Precision-focused)	Faster R-CNN	Highest precision-oriented composite score

Note: Model recommendations are derived from composite scores computed from Precision, Recall, and F1-score.

5) Sensitivity analysis

To further assess the robustness of the proposed framework, a sensitivity analysis was conducted by varying the Recall weight in the Security/Surveillance use case from 0.20 to 0.70. The remaining proportion of the weight was redistributed between Precision and F1-score to ensure that the total weighting remained normalized to one. This analysis was designed to capture the effect of shifting operational priorities, particularly the emphasis on minimizing missed detections.

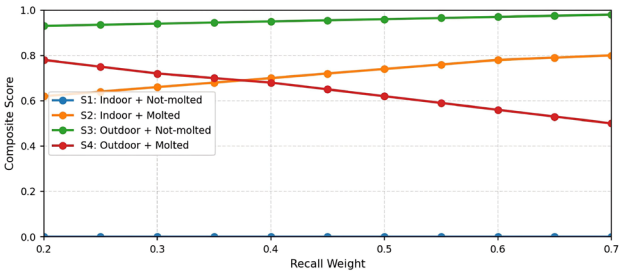

Fig. 5. Sensitivity analysis of composite scores in the security/surveillance use case under varying Recall weights (0.20–0.70).

As shown in Fig. 5, Scenario 3 (S3: Outdoor + Non-molted) consistently maintained the highest composite score across the entire range of Recall weights, thereby confirming its stability and robustness. In contrast, the relative positions of Scenario 2 (S2: Indoor + Molted) and Scenario 4 (S4: Outdoor + Molted) fluctuated depending on the Recall weight assigned. Specifically, S2 tended to gain an advantage under higher Recall weighting, while S4 performed better when Precision carried relatively more influence. These variations demonstrate that the proposed framework dynamically adapts to changes in use-case requirements while preserving the overall superiority of S3.

Scenario 3 (Outdoor + Non-molted) consistently achieved the highest composite score across all weighting configurations, demonstrating strong robustness to changes in Recall emphasis. In contrast, the relative rankings of Scenario 2 (Indoor + Molted) and Scenario 4 (Outdoor + Molted) shifted as the Recall weight increased, reflecting inherent trade-offs between Recall and other performance criteria.

6) Discussion of objectives

The findings of this study can be directly mapped to the three stated research objectives, as discussed below.

Objective 1: Development of a Framework. The study successfully established a generalizable metrics-based framework that integrates multi-dimensional performance indicators through a weighted decision-making process. By incorporating Precision, Recall, and F1-score into a unified composite index, the framework enables systematic and context-independent evaluation across diverse application domains.

Objective 2: Validation of Capability. The framework demonstrated empirical adaptability by producing context-specific recommendations, for instance, Scenario 2 (Indoor + Molted) was prioritized in the Security/Surveillance use case due to its higher Recall, whereas Scenario 4 (Outdoor + Molted) was more suitable for Quality Inspection because of its superior Precision. These outcomes confirm that the framework dynamically adjusts to operational priorities while maintaining robustness across scenarios.

Objective 3: Comparison with the Baseline. When compared with the single-metric baseline that relied solely on the F1-score, the proposed framework provided statistically validated differentiation, particularly between Scenarios 2 and 4. While the baseline approach ranked them similarly, the multi-criteria framework highlighted distinct trade-offs that are critical for decision support. This underscores the framework's added value in offering deeper insights beyond traditional single-metric evaluation.

B. Discussion

The results of this study provide new insights into the limitations of single-metric model selection and highlight the advantages of a multi-metric, MCDM-based framework for object detection model evaluation, unlike conventional approaches that prioritize accuracy or mAP [14, 15], In this study, the proposed framework evaluates multiple indicators, such as precision, recall, and F1-score, to expose trade-offs that may be obscured in single-dimensional assessments. The framework is extensible and can incorporate efficiency-related metrics for deployment-oriented evaluations.

A deeper examination of the precision-recall dynamic across the four scenarios clarifies the model's suitability for deployment. The divergence in S4, where precision was highest while recall dropped, indicates that the detector becomes conservative under visual ambiguity, which is likely due to confidence-based suppression of uncertain predictions. This behavior reduces false positives, but increases missed detections, which is

undesirable in safety-critical environments, such as hatchery monitoring and automated inspection. Conversely, the strong and balanced performance of S3 shows that consistent illumination enhances textural contrast and improves feature discrimination, making S3 more robust for general-purpose deployment. These findings reinforce the need for object detection evaluation to account not only for raw metric values, but also for the operational risks introduced by shifts in the precision–recall balance, particularly for visually dynamic biological specimens.

1) Precision–recall trade-offs across deployment contexts

A comparative analysis shows that each model occupies a distinct equilibrium in the precision–recall trade-off space. S2 is optimized for recall, which aligns with high-risk domains, such as medical diagnosis where missed detections pose serious consequences [41]. In contrast, S4 yields the highest precision, reflecting the requirements of industrial inspection systems that must minimize false alarms to ensure operational reliability [42]. Meanwhile, S3 demonstrates balanced performance across both precision and recall, consistent with the principle in multi-criteria decision-making that no model is universally optimal across all contexts and that the most suitable choice depends on application-specific priorities [39, 43]. These findings therefore reinforce that the "best" model is determined by the deployment setting rather than by any isolated performance metric.

2) Multi-Criteria Decision-Making (MCDM) integration

The ranking differences across scenarios illustrate the effectiveness of MCDM techniques such as Weighted Scoring and AHP [36, 37], enabling practitioner-controlled prioritization of precision, recall, latency, and resource cost. By combining decision science with computer vision, the framework transforms performance evaluation into structured decision support rather than raw benchmarking [43–46].

3) Implications for real-time and edge AI

Although latency and resource usage were not empirically measured in this experiment, their inclusion is important for model deployment under hardware constraints. Prior studies in edge AI and IoT show that lighter architectures and optimization techniques such as quantization and pruning [47, 48], as well as models like MobileNetV2 [49], can substantially reduce inference cost. Integrating such metrics into the proposed framework would enable recommendations based not only on accuracy, but also on latency and hardware feasibility, which are crucial for applications such as autonomous driving and industrial IoT [50–53].

4) Explainability, transparency, and trust

While explainability was not the primary target of this work, visualizations such as radar charts and sensitivity graphs clarify the rationale behind model rankings. This aligns with trends in Explainable AI (XAI) [54, 55] toward transparency in high-stakes AI adoption [56, 57], enabling domain experts without ML expertise to participate confidently in model-selection decisions [58].

C. Limitations and Future Work

Despite its contributions, the proposed framework has several limitations. First, the experimental scope included only four object detection models (YOLOv5, YOLOv7, SSD, and Faster R-CNN), excluding other widely adopted detectors such as RetinaNet and MobileNet-based SSD [1, 22]. Second, latency and computational efficiency were not empirically evaluated, leaving the suitability of the framework for embedded and edge-AI deployments unvalidated [9, 11]. Third, the weighting mechanism relied primarily on expert judgment, which is practical, but static and does not dynamically adapt to changing deployment conditions.

Future research could extend the framework in two complementary directions. From an application perspective, tailoring the weighting strategy to sector-specific requirements, such as precision-critical industrial inspection [42] or recall-critical medical diagnostics [41], which would enable more domain-aware decision support. From a methodological perspective, incorporating adaptive weighting mechanisms via reinforcement learning, Bayesian optimization, or multi-armed bandit algorithms may allow the framework to adjust criteria priorities in response to evolving environmental or operational constraints, transforming it into a self-adapting, intelligent decision-support system [59, 60].

Furthermore, the framework is inherently model-agnostic and can be scaled to emerging architectures. Extending the evaluation to YOLOv8 [21], YOLOv9, and transformer-based models such as DETR [61] and DINO [62] would broaden generalizability and demonstrate the ability of the decision mechanism to accommodate both single-stage and attention-based paradigms. Such expansion would also provide empirical insight into how innovations in speed, precision, and feature representation influence decision outcomes across real-world contexts.

Finally, it is important to emphasize that the contribution of this study lies not in comparing object detection models per se, but in operationalizing model selection as a multi-criteria decision-making process. Rather than determining which model achieves the highest accuracy, the proposed framework identifies the model that is most suitable for a given deployment scenario based on prioritized performance requirements. This distinction represents the conceptual novelty of the work and provides practical value for real-world AI deployment.

Future work will empirically measure inference latency (batch size = 1) and resource usage, including the number of parameters, FLOPs, and peak GPU memory consumption, under standardized hardware settings. These efficiency-oriented metrics will then be incorporated into the composite score to support real-time and resource-constrained deployments.

V. Conclusion

These insights highlight the practical significance of the proposed framework for real-world deployment, providing actionable guidance for model selection beyond traditional performance benchmarking. This study set out to address the critical limitations in conventional model selection for object detection, where evaluations are often narrowly determined by a single indicator such as mean mAP. Returning to the problem framed in the introduction, the research demonstrated that this reductionist approach fails to capture the multidimensional trade-offs inherent in deploying models across diverse operational contexts. By proposing a Metrics-Based Framework that integrates multiple performance criteria and decision-making techniques, this work contributes to the development of a more comprehensive, transparent, and context-sensitive paradigm for model evaluation.

The contributions of this study can be articulated along two interrelated dimensions.

- **Theoretical Contribution:** This work bridges the domains of computer vision and multi-criteria decision-making (MCDM), showing how weighted scoring, AHP, and related methods can systematically reconcile trade-offs among precision, recall, latency, and resource usage. While prior studies have focused predominantly on accuracy benchmarks, the present framework underscores that no single model is universally optimal. Instead, model suitability is conditional upon domain-specific requirements and stakeholder priorities. The conceptual integration advanced here introduces a new lens for the field, highlighting how decision science can enrich the theoretical understanding of object detection evaluation.
- **Practical Contribution:** On the applied side, the framework translates complex performance data into actionable insights that practitioners in security, healthcare, transportation, and manufacturing can use directly. By presenting composite scores, radar charts, heatmaps, and the use of case–specific rankings, the framework enables stakeholders to identify not just which model performs "best" in aggregate, but which model is most appropriate for their operational needs. This is particularly valuable in real-world deployments, where trade-offs between recall and latency or between precision and computational cost can have significant consequences for safety, efficiency, or cost-effectiveness.

Beyond these contributions, the study offers important implications for future research and practice. First, it demonstrates that systematic, decision-oriented frameworks can democratize model selection, supporting both experts and non-experts in making informed deployment choices. Second, it provides a foundation for extending evaluation pipelines toward adaptive and risk-aware decision support, where reinforcement learning, Bayesian optimization, or other adaptive weighting mechanisms are able to dynamically adjust priorities based on changing environments or data distributions. Third, it invites dialogue between the computer vision and operations research communities, suggesting that interdisciplinary collaboration can produce tools that are both methodologically rigorous and practically relevant.

Overall, the findings reaffirm the central thesis of this research: that effective object detection requires not only ongoing advances in model architecture, but also structured evaluation frameworks that translate technical results into context-aware recommendations. By closing this circle, the study underscores the value of linking theoretical innovation with practical utility, paving the way for future work on adaptive, domain-specific, and risk-sensitive decision support systems in computer vision.

Although the current evaluation focused on four representative detectors, the framework is model-agnostic and can be extended to newer architectures as long as measurable metrics are available. This generalizability opens opportunities to evaluate emerging paradigms such as the YOLOv8 series for edge AI and DETR-based transformer architectures for complex scene understanding. Future work will prioritize incorporating these architectures to further verify the scalability and long-term relevance of the framework, and will also integrate empirical latency and computational cost measurements to enhance its practicality for deployment in real-time and edge-AI environments.

Conflict of Interest

The authors declare that they have no known financial, professional, or personal conflicts of interest that could have influenced the work reported in this manuscript.

Author Contributions

Sudasawan Ngammongkolwong: Research conceptualization, research methodology design, data analysis, development of the conceptual framework, manuscript writing, and manuscript revision. Rungtiwa Saosing: Statistical validation, refinement of experimental procedures, and critical review of the manuscript. All authors had approved the final version.

Funding

This publication was financially supported by Southeast Bangkok University, Thailand.

Acknowledgment

The authors would like to thank Southeast Bangkok University and Rajamangala University of Technology Krungthep for the academic support and resources that facilitated this research.

References

[1] W. Liu, D. Anguelov, D. Erhan *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 580–587.

[3] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019. doi: 10.1109/ACCESS.2019.2939201

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2016.

[7] T. Y. Lin, P. Dollár, R. Girshick *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[8] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023. doi: 10.1109/JPROC.2023.3238524

[9] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7310–7311.

[10] J. C. Nascimento and J. S. Marques, "Performance evaluation of object detection algorithms for video surveillance," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 761–774, Aug. 2006. doi: 10.1109/TMM.2006.876287

[11] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv Preprint, arXiv:2004.10934, 2020.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2001, pp. 511–518.

[13] X. Wu, D. Sahoo, and S. C. H. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, 2020.

[14] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1440–1448. doi: 10.1109/ICCV.2015.169

[15] D. Cantero, I. Esnaola-Gonzalez, J. Miguel-Alonso, and E. Jauregi, "Benchmarking object detection deep learning models in embedded devices," *Sensors*, vol. 22, no. 11, 4205, 2022. doi: 10.3390/s22114205

[16] M. Everingham *et al.*, "The pascal Visual Object Classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[17] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv Preprint, arXiv:1804.02767, 2018.

[18] C. Wang *et al.*, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7464–7475.

[19] G. Jocher *et al.* (2020). YOLOv5 release documentation. *GitHub Repository*. [Online]. Available: https://github.com/ultralytics/yolov5

[20] G. Jocher, A. Chaurasia, J. Qiu *et al.*, "YOLOv8: Ultralytics' next-generation real-time object detector," Ultralytics Technical Report, 2023.

[21] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10781–10790.

[22] T. Y. Lin *et al.*, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988.

[23] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1086–1095.

[24] W. Wei, Y. Cheng, J. He *et al.*, "A review of small object detection based on deep learning," *Neural Comput. & Applic.*, vol. 36, pp. 6283–6303, 2024. https://doi.org/10.1007/s00521-024-09422-6

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," arXiv Preprint, arXiv:1405.0312, 2014.

[26] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9.

[27] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," arXiv Preprint, arXiv:2010.16061, 2020.

[28] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[29] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 233–240.

[30] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge, U.K.: Cambridge Univ. Press, 2008.

[31] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7263–7271.

[32] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[33] T. Nowak, K. Ćwian, and P. Skrzypczyński, "Real-time detection of non-stationary objects using intensity data in automotive LiDAR SLAM," *Sensors*, vol. 21, no. 20, 6781, 2021. doi: 10.3390/s21206781

[34] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," arXiv Preprint, arXiv:1712.04621, 2017.

[35] T. Czimmermann, G. Ciuti, M. Milazzo, M. Chiurazzi, S. Roccella, C. M. Oddo, and P. Dario, "Visual-based defect detection and classification approaches for industrial applications—A survey," *Sensors*, vol. 20, no. 5, 1459, 2020. doi: 10.3390/s20051459

[36] C. L. Hwang and K. Yoon, *Multiple Attribute Decision Making: Methods and Applications*, New York, NY, USA: Springer-Verlag, 1981.

[37] T. Saaty, "Analytic hierarchy process," in *Encyclopedia of Operations Research and Management Science*, Boston, MA, USA: Springer, 1980.

[38] V. Belton and T. Stewart, *Multiple Criteria Decision Analysis: An Integrated Approach*, New York, NY, USA: Springer, 2002.

[39] S. Opricovic and G. H. Tzeng, "Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS," *Eur. J. Oper. Res.*, vol. 156, no. 2, pp. 445–455, 2004. doi: 10.1016/S0377-2217(03)00020-1

[40] K. Govindan and M. B. Jepsen, "ELECTRE: A comprehensive literature review on methodologies and applications," *Eur. J. Oper. Res.*, vol. 250, no. 1, pp. 1–29, 2016.

[41] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[42] J. Xu, M. Kovatsch, D. Mattern *et al.*, "A review on AI for smart manufacturing: Deep learning challenges and solutions," *Applied Sciences*, vol. 12, no. 16, 8239, 2022. doi: 10.3390/app12168239

[43] J. Figueira, S. Greco, and M. Ehrgott, *Multiple Criteria Decision Analysis: State of the Art Surveys*, New York, NY, USA: Springer, 2005.

[44] S. Minaee *et al.*, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, 2021.

[45] E. K. Zavadskas, Z. Turskis, and J. Tamošaitienė, "Selection of construction enterprises management strategy based on the SWOT and multi-criteria analysis," *Archives of Civil and Mechanical Engineering*, vol. 11, no. 4, pp. 1063–1082, 2011. doi: 10.1016/S1644-9665(12)60096-X

[46] J. Lee, B. Bagheri, and H.-A. Kao, "A cyber-physical systems architecture for Industry 4.0-based manufacturing systems," *Manufacturing Letters*, vol. 3, pp. 18–23, 2015. doi: 10.1016/j.mfglet.2014.12.001.

[47] M. Chen, Y. Hao, K. Hwang *et al.*,"Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017. doi: 10.1109/ACCESS.2017.2694446

[48] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2016.

[49] B. Jacob *et al*., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2704–2713. doi: 10.1109/CVPR.2018.00286

[50] N. D. Lane *et al.*, "Deepx: A software accelerator for low-power deep learning inference on mobile devices," in *Proc. International Conference on Information Processing in Sensor Networks (IPSN)*, 2016, pp. 1–12. doi: 10.1109/IPSN.2016.7460664

[51] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474
[52] K. Duan, S. Bai, L. Xie *et al.*, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 6568–6577. doi: 10.1109/ICCV.2019.00667
[53] A. Geiger *et al.*, "Vision meets robotics: The KITTI dataset," *Int. J. Robotics Res.*, vol. 32, no. 11, pp. 1231–1237, 2013. doi: 10.1177/0278364913491297
[54] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.
[55] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.
[56] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv Preprint, arXiv:1702.08608, 2017.
[57] W. Samek, T. Wiegand, and K. R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," arXiv Preprint, arXiv:1708.08296, 2017.
[58] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, North Carolina, USA: Leanpub, 2022.
[59] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[60] K. Mokhtari and A. R. Wagner, "Don't get yourself into trouble! Risk-aware decision-making for autonomous vehicles," arXiv Preprint, arXiv:2106.04625, 2021.
[61] N. Carion, F. Massa, G. Synnaeve *et al.*, "End-to-end object detection with transformers," in *Proc. European Conf. Computer Vision (ECCV)*, 2020, pp. 213–229.
[62] H. Zhang, F. Li, S. Liu *et al.*, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," arXiv Preprint, arXiv:2203.03605, 2022.