

Comparative Analysis of YOLOv5, YOLOv8, and YOLOv11 for Military Aircraft Detection on Aerial Imagery Supporting Airspace Surveillance

Fajar Sidik Suganda^{1*}, Asep Adang Supriyadi²

^{1,2}Sensing Technology Study Program, Republic of Indonesia Defense University, Sentul, West Java, Indonesia
E-mail: ^{1*}fajarsidiksugandaa@gmail.com, ²aadangsupriyadi@gmail.com

(Received: 28 Apr 2026, revised: 12 May 2026, accepted: 13 May 2026)

Abstract

Automated military aircraft detection from aerial imagery is a key enabler for modern Intelligence, Surveillance, and Reconnaissance (ISR) operations, particularly for nations with vast airspace such as Indonesia. While the You Only Look Once (YOLO) family has become a de facto standard for such tasks, the literature lacks a controlled, head-to-head comparison of its recent generations on a standard military aircraft benchmark, which leaves architecture selection for ISR systems largely guided by version recency rather than empirical evidence. This study addresses that gap by benchmarking three YOLO generations YOLOv5s, YOLOv8s, and YOLOv11s on the MAR20 dataset (3,842 images, 20 classes, 22,341 instances) under identical training conditions, and by analyzing the results along three axes: aggregate accuracy, per-class behavior, and accuracy–parameter trade-off. Counter to the assumption that newer is necessarily better, the oldest architecture in the comparison, YOLOv5s, achieves the highest Precision (0.985) and F1-Score (0.978) with the smallest parameter count (7.2M), while YOLOv11s, the most recent iteration, leads only on Recall (0.972) and YOLOv8s offers the most balanced profile; all three exceed mAP@0.5 of 0.988. Per-class analysis further shows that detection difficulty is driven by inter-class visual similarity rather than class imbalance, with visually similar fighter aircraft (F-22, F-16, F-15) forming a distinct hard-class cluster (mean AP 0.705–0.788) despite adequate training samples. Based on these findings we propose an operationally-mapped model-selection guide YOLOv5s as a strong candidate for edge-constrained ISR platforms, YOLOv8s for balanced ground-station workloads, and YOLOv11s for recall-critical wide-area surveillance while explicitly noting that the recommendations rest on accuracy–parameter evidence rather than on-device latency measurements, and that the single-seed, small-variant setting bounds the generality of the ranking. Validation through repeated runs with statistical testing, profiling on representative edge hardware, and evaluation on operational imagery remain important directions for future work.

Keywords: YOLO, Object Detection, Military Aircraft, Aerial Imagery, Airspace Surveillance.

I. INTRODUCTION

Airspace sovereignty is a fundamental pillar of modern national defense systems, particularly for Indonesia, which possesses a vast airspace territory with complex archipelagic geography and is traversed by three Indonesian Archipelagic Sea Lanes (ALKI) that make it a strategic point in international aviation [1]. In preserving such sovereignty, Intelligence, Surveillance, and Reconnaissance (ISR) operations play a strategic role in continuously detecting, identifying, and monitoring aerial object movements, as mandated by Law No. 34 of 2004 on the Indonesian National Armed Forces, which grants the Indonesian Air Force the authority to safeguard the country's airspace [2]. Advances in Unmanned Aerial Vehicle (UAV) technology and high-resolution satellite imagery have significantly expanded the aerial data acquisition capabilities

that can be leveraged for defense and reconnaissance purposes [3]. In parallel, the progress of deep learning within the computer vision domain has introduced object detection algorithms capable of delivering accurate and real-time performance, with the You Only Look Once (YOLO) family emerging as one of the most widely adopted one-stage detector approaches across numerous scientific studies [4] [5]. The evolution of this algorithm began with YOLOv1 in 2015 and continues to YOLOv11 released by Ultralytics in September 2024, with each iteration introducing significant architectural refinements, ranging from anchor-free detection in YOLOv8 to the C3k2, SPPF, and C2PSA modules in YOLOv11 that strengthen feature extraction on complex imagery [6]. The integration of aerial image acquisition platforms with YOLO algorithms opens substantial opportunities for developing



automated airspace surveillance systems that support both tactical and strategic decision-making processes [7]

Although YOLO algorithms have been extensively applied across various domains such as cultural pattern detection, script recognition, and facial expression analysis [8] [9], their implementation for military aircraft detection in aerial imagery remains highly limited, particularly within the Indonesian defense context. Airspace violations by foreign aircraft continue to occur repeatedly, as recorded in the incidents involving U.S. F-18 Hornet military aircraft over Bawean Island in 2003, Japan Navy P3C Orion military aircraft in 2019, and the forced landing of a Boeing 777 cargo aircraft at Hang Nadim Airport in January 2019, demonstrating that conventional radar-based early detection systems still have limitations due to radar gaps, particularly around Natuna and Kalimantan. Most existing studies on aerial imagery have focused on civilian objects such as ground vehicles, buildings, or general infrastructure, and have yet to comprehensively investigate strategic military targets [10].

The study by Elgamily et al. which optimized YOLOv7 and YOLOv8 using metaheuristic algorithms on remote sensing datasets, only covered generic objects such as vehicles and buildings, without focusing on the specific classification of military aircraft. Similarly, the study by Cinarer et al, which compared U-Net and YOLO models for aircraft detection on satellite imagery, has not included the YOLOv11 architecture, which is the latest version. Comparative studies simultaneously encompassing the three most recent YOLO architectures (YOLOv5, YOLOv8, and YOLOv11) for military aircraft detection are still scarce, leaving ISR system developers without strong empirical references for selecting the most suitable architecture [11]. The technical challenges of military aircraft detection in aerial imagery are also distinctive, including relatively small object sizes (small object detection), high visual similarity among aircraft classes, and intra-class variations caused by differences in viewpoint, illumination, and atmospheric conditions [12]. Furthermore, per-class performance evaluation for specific military aircraft types such as fighters, bombers, and transport aircraft has received inadequate attention in the national literature, despite its critical importance in assessing the operational feasibility of such systems [13].

This research is essential because the modernization of Indonesia's air defense system demands the adoption of accurate and efficient artificial intelligence technologies to support early detection of airspace violations, while the lack of empirical references regarding the performance of recent YOLO architectures on military aircraft objects may hinder both technical and strategic decision-making by defense stakeholders. To address this gap, this study offers a novel approach through a systematic comparative benchmarking of three state-of-the-art YOLO architectures (YOLOv5, YOLOv8, and YOLOv11) on the MAR20 dataset, which is a standard dataset for military aircraft recognition from remote sensing imagery, employing comprehensive evaluation metrics including mean Average Precision (mAP), precision, recall, F1-score, and per-class performance analysis. This research aims to produce empirical recommendations

regarding the most optimal YOLO architecture for implementation in military aircraft detection systems based on aerial imagery, thereby supporting the development of automated airspace surveillance systems within the framework of strengthening Indonesia's national defense. The main contributions of this research can be summarized as follows: (1) presenting, to the best of our knowledge, one of the first comparative analyses in the national literature that simultaneously benchmarks YOLOv5, YOLOv8, and YOLOv11 on the MAR20 military aircraft dataset; (2) providing detailed per-class performance evaluation for 20 types of military aircraft, which serves as a practical reference for operational implementation; and (3) offering technical recommendations relevant to the Indonesian airspace surveillance context. The remainder of this paper is organized as follows: Section 2 presents a literature review on YOLO architectures and aerial object detection; Section 3 describes the research methodology including the dataset, experimental setup, and evaluation metrics; Section 4 presents the experimental results and discussion; and Section 5 provides conclusions and directions for future research.

II. LITERATURE REVIEW

A. Airspace Surveillance and ISR Operations

Airspace surveillance represents an integral component of national defense systems aimed at detecting, identifying, and assessing the movement of aerial objects within a country's sovereign territory. Within the operational framework of modern defense, such activities are conducted through the Intelligence, Surveillance, and Reconnaissance (ISR) concept, which integrates three key functions: intelligence as the collection and analysis of strategic information, surveillance as continuous monitoring of specific targets, and reconnaissance as active exploration of operational areas. Contemporary ISR systems integrate diverse sensor platforms, ranging from ground-based radar, reconnaissance aircraft, Unmanned Aerial Vehicles (UAVs), to high-resolution satellite imagery, in order to produce comprehensive situational awareness [14]. For Indonesia, implementing ISR presents specific challenges due to its extensive archipelagic geography, the existence of radar coverage gaps in several strategic regions, and the limitations of conventional defense assets, all of which drive the need for new artificial intelligence-based technological approaches to complement existing detection systems [15]. The utilization of machine learning technology on aerial imagery data is regarded as a promising solution because it can produce automated detection capabilities that are fast, consistent, and operable on a continuous basis at a scale that is difficult for human operators to achieve [16].

B. Deep Learning-Based Object Detection

Methodologically, deep learning-based object detection approaches can be categorized into two main groups: two-stage detectors such as Faster R-CNN and Mask R-CNN, which separate the region proposal stage from the classification stage, and one-stage detectors such as YOLO,

SSD, and RetinaNet, which perform bounding box and class predictions simultaneously in a single forward pass. The one-stage approach generally delivers higher inference speed, making it more suitable for real-time applications, while two-stage approaches tend to yield slightly higher accuracy at the cost of longer computation time. In the domain of aerial and remote sensing imagery, deep learning-based approaches have proven superior to conventional methods relying on hand-crafted features, particularly in handling complex variations of scale, orientation, and illumination. Hussain [17] in his comprehensive review concluded that the YOLO family has become the de facto standard for real-time object detection applications due to its optimal balance between inference speed, detection accuracy, and ease of implementation across various computing platforms, making it the primary choice for researchers and industry practitioners across various application domains.

C. Evolution of YOLO Architectures

YOLOv5, released by Ultralytics in 2020, introduced the CSPDarknet53 architecture in the backbone, which adopts the Cross Stage Partial (CSP) strategy to split feature maps into two paths, effectively reducing computational redundancy while improving gradient flow during training. This architecture is complemented by the Spatial Pyramid Pooling Fast (SPPF) module in the neck section, which enables efficient multi-scale feature extraction, and utilizes the Path Aggregation Network (PANet) for multi-level feature fusion from the backbone. YOLOv5 still adopts an anchor-based detection approach with anchor boxes optimized through a k-means clustering algorithm, which has proven to provide competitive performance on various popular benchmarks. YOLOv8, released by Ultralytics in January 2023, brought several significant architectural changes, including the transition to an anchor-free detection approach that simplifies the Non-Maximum Suppression (NMS) process during post-processing, the replacement of the C3 module in YOLOv5 with the C2f module which provides richer feature representation by combining outputs from all bottleneck layers, and the adoption of a decoupled head that separates the classification pathway from bounding box regression [18]. Additionally, YOLOv8 introduces Distribution Focal Loss (DFL), which has been shown to provide smoother and more informative localization gradients for dense object detection tasks. YOLOv11, the latest iteration introduced at YOLO Vision 2024, presents three key innovations: the C3k2 module that replaces C2f with a lighter 3×3 kernel-based bottleneck, an enhanced SPPF module for multi-scale feature extraction, and the Convolutional block with Parallel Spatial Attention (C2PSA) module that strengthens spatial attention capabilities toward relevant object regions in high-resolution images. Various recent architectural modification studies have also demonstrated that YOLOv8 and YOLOv11 have flexible optimization spaces, such as integration of attention mechanisms [19], modified feature pyramid networks [20], and optimized loss functions such as Inner-SIoU.

D. AI-Based ISR Systems in Other Countries: Comparison of the USA, China, and Israel

The application of artificial intelligence to Intelligence, Surveillance, and Reconnaissance (ISR) systems has become a strategic development focus in various advanced nations, with three countries standing out most prominently: the United States, China, and Israel. The United States developed Project Maven, which began in 2017 under the U.S. Department of Defense, as an AI initiative to automate the processing of full-motion video imagery from reconnaissance aircraft and satellites using computer vision algorithms capable of detecting and classifying tactical objects in real-time, with the ability to generate thousands of target recommendations per hour, and has been widely adopted by various U.S. combatant commands [21]. China, through its People's Liberation Army (PLA), develops the concept of intelligentized warfare, which integrates AI with Command, Control, Communications, Computers, Intelligence, Surveillance, Reconnaissance, and Targeting (C5ISRT) systems, with substantial investment in automated target recognition, intelligent autonomous vehicles, and deep learning-based recognition systems integrated into UAVs and missile systems [22]. Israel, as a pioneer in the defense AI domain, has developed systems such as Fire Weaver from Rafael Advanced Defense Systems, which utilizes computer vision to connect sensors with shooters in a networked manner, as well as Iron Dome, which uses AI to improve missile interception accuracy to over 90%, complemented by the Habsora system for operational target analysis and recommendation [23]. Comparison of these three countries reveals that although each country's approach differs, all emphasize computer vision-based deep learning as a fundamental component of modern ISR systems, providing an important perspective that the development of YOLO-based military aircraft detection modules can be a rational first step to building Indonesia's national ISR capacity.

III. RESEARCH METHODOLOGY

A. Research Design

This research employs a quantitative research method with a comparative experimental approach (comparative experimental study), designed to systematically compare the performance of three recent YOLO architectures on the task of military aircraft detection. This design was selected because it enables objective evaluation based on measurable quantitative metrics that can be reproduced by other researchers, in line with benchmarking practices commonly applied in deep learning-based military aircraft detection studies [24] [25]. Each model was trained and evaluated under identical experimental conditions, including the dataset, data split partitioning, hyperparameters, and computing environment, so that performance differences produced can be primarily attributed to architectural differences between models, while acknowledging that stochastic factors such as weight initialization and data shuffling may also contribute and were not isolated through repeated runs in this study. The research

workflow consists of five main stages: dataset acquisition, preprocessing and data augmentation, model training, performance evaluation, and comparative analysis of the results. This controlled benchmarking approach aligns with standard practices in the computer vision community to ensure internal validity and comparability of results with previous studies [26].

B. Research Material (Dataset)

The primary material used in this research is the MAR20 (Military Aircraft Recognition Dataset), which is one of the largest publicly available datasets for military aircraft recognition from remote sensing imagery [27] [28]. The MAR20 dataset consists of 3,842 high-resolution images collected from 60 military airbases across various countries through Google Earth, with a total of 22,341 annotated aircraft instances distributed across 20 different military aircraft classes such as SU-35, TU-160, TU-22, TU-95, SU-34, and SU-24. Each object instance in the dataset is equipped with two annotation formats, namely horizontal bounding boxes and oriented bounding boxes; however, this research uses only the horizontal bounding box annotations as they are compatible with standard YOLO architectures. The dataset is divided into three subsets with a ratio of 70% for training (2,689 images), 15% for validation (577 images), and 15% for testing (576 images), performed using a stratified split to maintain balanced class distribution proportions across each subset and to mitigate the risk of data leakage between training and testing sets, which has been shown to substantially inflate reported performance when overlooked in deep learning evaluations [29]. The detailed instance distribution across the 20 aircraft classes is presented in Table 1 and visualized in Figure 1 and Figure 2.

Table 1. Instance Distribution of the 20 Military Aircraft Classes in the MAR20 Dataset.

Code	Class Name	Instance Count	Image Count	% Inst	% Img
A1	SU-35	1646	350	7.37	9.11
A2	C-130	1729	366	7.74	9.53
A3	C-17	1176	314	5.26	8.17
A4	C-5	642	147	2.87	3.83
A5	F-16	1262	288	5.65	7.5
A6	TU-160	441	108	1.97	2.81
A7	E-3	680	190	3.04	4.95
A8	B-52	944	205	4.23	5.34
A9	P-3C	1086	223	4.86	5.8
A10	B-1B	924	194	4.14	5.05
A11	E-8	507	186	2.27	4.84
A12	TU-22	702	224	3.14	5.83
A13	F-15	1652	277	7.39	7.21
A14	KC-135	1778	451	7.96	11.74
A15	F-22	618	145	2.77	3.77
A16	FA-18	2632	297	11.78	7.73
A17	TU-95	1397	376	6.25	9.79
A18	KC-10	308	101	1.38	2.63
A19	SU-34	1236	375	5.53	9.76

A20	SU-24	981	255	4.39	6.64
TOTAL		22341	3842	100	100

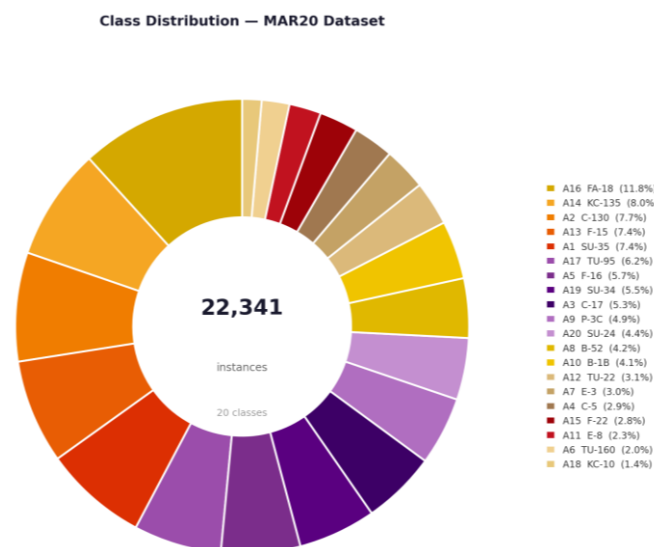


Figure 1. Visualization of the Class Distribution in the MAR20 Dataset.

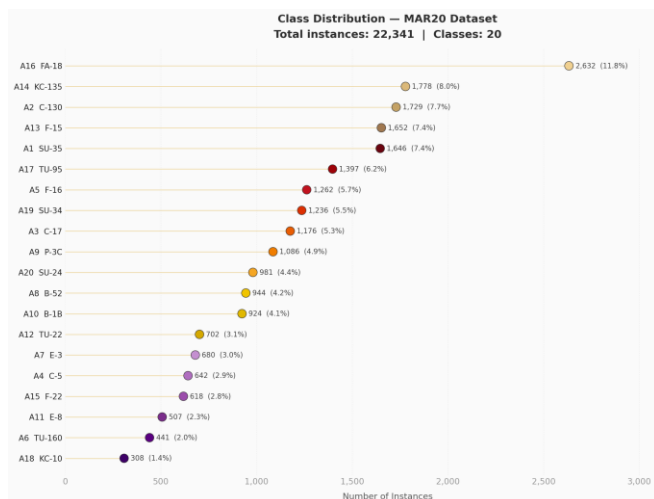


Figure 2. Instance Count per Class in Descending Order.

Analysis of the class distribution reveals a moderate class imbalance within the MAR20 dataset, with the number of instances per class ranging from 308 (A18 KC-10) to 2,632 (A16 FA-18), yielding an imbalance ratio of approximately 8.5:1 between the largest and smallest classes. The top three classes FA-18 (11.78%), KC-135 (7.96%), and C-130 (7.74%) collectively account for nearly 28% of all instances, while the bottom three classes KC-10 (1.38%), TU-160 (1.97%), and E-8 (2.27%) together represent only 5.62% of the dataset. Notably, several pairs of aircraft with high visual similarity exhibit highly skewed distributions, including KC-135 (A14) versus KC-10 (A18) as refueling tankers with a ratio of 5.8:1, TU-95 (A17) versus TU-160 (A6) as Russian strategic bombers with a ratio of 3.2:1, and E-3 (A7) versus E-8 (A11) as Boeing-platform surveillance aircraft with a ratio of 1.3:1. This imbalance, combined with inter-class visual similarity, is

expected to present fine-grained classification challenges during model training, particularly affecting detection performance on minority classes within each visually-similar pair.

C. Research Instruments and Procedure

The software instruments used in this research include the Python programming language version 3.10, the PyTorch deep learning framework version 2.0, and the Ultralytics library for the consistent implementation of YOLOv5, YOLOv8, and YOLOv11. The Roboflow platform is used for dataset management and application of data augmentation, while the training process is conducted on a Google Colab Pro cloud computing environment equipped with an NVIDIA Tesla T4 GPU with 16 GB of memory. Dataset preprocessing includes resizing all images to dimensions of 640×640 pixels, normalizing pixel values to the range [0,1], and converting annotation formats from MAR20 to YOLO format (normalized center x, center y, width, height). Data augmentation applied during training includes mosaic augmentation, which was first introduced in YOLOv4 and has been proven effective in enhancing small object detection capabilities, random horizontal flip, HSV color jitter, random scaling in the range of 0.5 - 1.5, and mixup, which collectively aim to improve model generalization against variations in object appearance [30]. The training procedure is conducted with identical configurations for the three models: 100 epochs, a batch size of 16, an input image size of 640×640, the SGD optimizer with an initial learning rate of 0.01 and momentum of 0.937, and weight decay of 0.0005 [31], with a warmup strategy in the first three epochs and a cosine learning rate schedule to gradually reduce the learning rate.

D. Data Analysis and Evaluation

Data analysis is performed quantitatively using standard evaluation metrics for object detection tasks, namely mean Average Precision (mAP@0.5), mean Average Precision across the IoU threshold range of 0.5 to 0.95 (mAP@0.5:0.95), Precision, Recall, and F1-score, which are calculated on the testing subset after the training process is completed [32]. In addition to aggregate metrics, per-class Average Precision analysis is conducted to assess the model's performance on each type of military aircraft, which is important for identifying classes that are difficult to detect due to visual similarity or imbalanced sample counts between classes. Confusion matrices are constructed for each model to visualize classification error patterns, particularly for class pairs with high confusion rates. Comparison between models is conducted directly on all of these metrics, supplemented by visual analysis in the form of training loss curves, validation mAP curves, and sample detection results on representative test samples for each architecture. The analysis results are then interpreted to produce recommendations for the most optimal YOLO architecture for military aircraft detection on aerial imagery in the context of airspace surveillance systems.

IV. RESULTS

A. Comparison of Main Metrics Across the Three Models

The training and evaluation results of the three YOLO architectures on the MAR20 testing set are summarized in Table 2. All three models demonstrated strong detection performance with mAP@0.5 values exceeding 0.988, confirming the effectiveness of YOLO architectures for military aircraft recognition on aerial imagery. A closer examination of the per-metric results reveals that each architecture exhibits its own distinct strengths: YOLOv5s achieved the best overall performance with mAP@0.5 = 0.991, Precision = 0.985, and F1-Score = 0.978, despite having the smallest parameter count (7.2M); YOLOv8s reached an equivalent mAP@0.5 but at 56% higher model complexity (11.2M parameters); while YOLOv11s recorded the highest Recall (0.972), indicating a more comprehensive detection capability across aircraft instances. The mAP@0.5:0.95 values across the three models are remarkably consistent (0.819–0.821), demonstrating stable bounding-box localization quality at stricter IoU thresholds. Across the three models, the ranking on Precision, Recall, F1-Score, and parameter count is summarized in Table 2; interpretation of these rankings and their implications is deferred to the Discussion (Section V).

Table 2. Overall Performance Comparison of Three YOLO Architectures on the MAR20 Testing Set.

Model	Precision	Recall	F1	mAP@0.5	mAP@0.5:0.95
YOLOv5s	0.985	0.971	0.9779	0.991	0.8209
YOLOv8s	0.9815	0.9662	0.9738	0.991	0.819
YOLOv11s	0.9673	0.9724	0.9698	0.988	0.8201

B. Training Loss and Validation mAP Curve Analysis

The training loss and validation mAP@0.5 curves over 100 epochs are presented in Figure 3, illustrating the convergence dynamics of the three models. All three architectures exhibit rapid loss reduction within the first 20 epochs, decreasing from approximately 1.30 to 0.85, followed by a more gradual descent toward the final loss values of 0.89 (YOLOv5s), 0.84 (YOLOv8s), and 0.89 (YOLOv11s). The validation mAP@0.5 curves show that all three models reach approximately 0.95 by epoch 20 and stabilize above 0.97 after epoch 40, with continued marginal improvements through to convergence at the end of training. A brief instability is observed during the early training phase (epochs 1–10), particularly for YOLOv11s, which experiences a temporary mAP fluctuation around epoch 6 before recovering rapidly. After this transient, all three curves converge smoothly without further perturbation.

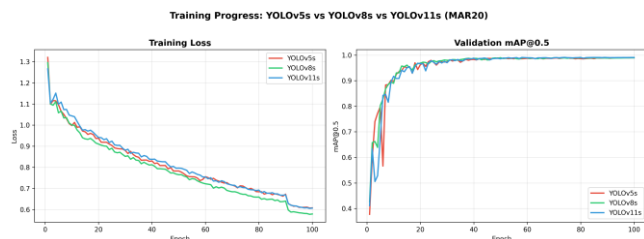


Figure 3. Training Loss and Validation mAP@0.5 Across 100 Epochs.

C. Confusion Matrix Analysis

The normalized confusion matrices for the three models on the test set are presented in Figure 4, with diagonal elements representing correct classifications and off-diagonal elements representing inter-class misclassifications. All three models exhibit strongly diagonal-dominant patterns, with diagonal values ranging from 0.93 to 1.00, indicating high classification consistency across the 20 aircraft classes. Inter-class misclassifications between aircraft types are minimal (≤ 0.05 in most cases), confirming that the three models effectively distinguish among different aircraft categories. The most notable observation is that the primary detection challenge lies in the aircraft-background separation rather than in fine-grained classification among aircraft classes, with values in the rightmost "background" column ranging from 0.02 to 0.15. Higher background-confusion rates are observed for several classes including A15_F-22 (0.06–0.12), A1_SU-35 (0.11–0.12), A11_E-8 (0.08–0.15), A13_F-15 (0.08–0.14), A14_KC-135 (0.06–0.14), and A19_SU-34 (0.11–0.14). Among the three models, YOLOv11s records a slightly higher overall Recall (0.972) than YOLOv5s and YOLOv8s.

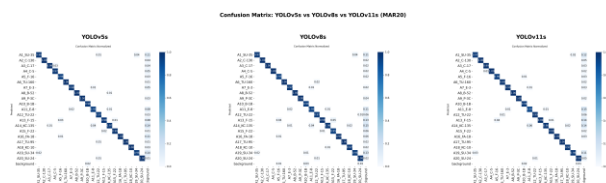


Figure 4. Confusion Matrices for YOLOv5s, YOLOv8s, and YOLOv11s.

D. Analysis of Classes with the Lowest Detection Rates

The per-class Average Precision (AP) analysis (Table 3) shows meaningful variability across the 20 aircraft classes, with mean AP across the three models ranging from 0.705 (A15_F-22) to 0.903 (A17_TU-95). The three classes with the lowest detection performance are A15_F-22 (mean AP = 0.705), A5_F-16 (mean AP = 0.770), and A13_F-15 (mean AP = 0.788), all fighter aircraft of comparable physical dimensions. Classes with among the smallest sample counts KC-10 (A18, 308 instances) and TU-160 (A6, 441 instances) produced AP values of 0.822 and 0.823 respectively, placing them in the middle of the distribution rather than at the bottom. At the upper end, the bomber and transport aircraft A17_TU-95, A8_B-52, and A12_TU-22 each achieved AP above 0.85. Interpretation of these per-class differences in particular the

role of inter-class visual similarity relative to sample distribution is presented in Section V.

Table 3. Per-Class Average Precision

Class	YOLOv5s	YOLOv8s	YOLOv11s
A1_SU-35	0.8481	0.8577	0.8473
A2_C-130	0.8126	0.8248	0.8173
A3_C-17	0.8305	0.8129	0.8455
A4_C-5	0.8402	0.7953	0.8127
A5_F-16	0.7787	0.7618	0.7704
A6_TU-160	0.819	0.8217	0.8296
A7_E-3	0.8488	0.8445	0.8705
A8_B-52	0.8618	0.8688	0.8648
A9_P-3C	0.7919	0.7981	0.7956
A10_B-1B	0.8114	0.804	0.8149
A11_E-8	0.8255	0.8255	0.7824
A12_TU-22	0.8665	0.8701	0.8406
A13_F-15	0.7875	0.79	0.7853
A14_KC-135	0.8147	0.8275	0.8305
A15_F-22	0.701	0.6928	0.7211
A16_FA-18	0.8208	0.8123	0.8199
A17_TU-95	0.8984	0.9039	0.906
A18_KC-10	0.8141	0.8228	0.8285
A19_SU-34	0.8176	0.8165	0.8058
A20_SU-24	0.829	0.8285	0.813

E. Qualitative Detection Performance

Figure 5 presents qualitative detection outputs from the three models on representative test samples, illustrating bounding-box localization and class assignment on aircraft at varying scales and orientations within the airbase scenes. Across the displayed samples, the three architectures produce visually similar detections on well-separated targets, while differences in bounding-box tightness and confidence scores are most apparent on closely-spaced or partially-occluded instances. Discussion of what these qualitative patterns imply for model selection and deployment is presented in Section V.



Figure 5. Detection Results on Three Representative Test Samples

V. DISCUSSION

One of the most fundamental contributions of this study is the provision of a systematic comparative analysis of YOLOv5, YOLOv8, and YOLOv11 on the MAR20 dataset something that was previously unavailable in the literature. By applying identical hyperparameters and training procedures across the three architectures, performance differences that emerge can be primarily attributed to architectural design rather than methodological bias, subject to the caveat that single-run results do not separate architectural effects from stochastic training variance. This contribution enriches the body of knowledge on YOLO benchmarking for military remote sensing imagery, which had previously been dominated by partial evaluations or single-model implementations [33].

A more compelling finding emerges when the experimental results challenge the conventional assumption that newer architectures always deliver superior performance. YOLOv5s, the oldest architecture in this comparison, achieves the highest Precision (0.985) and the highest F1-Score (0.978) with the smallest parameter count (7.2M). Meanwhile, YOLOv11s the most recent iteration excels only on the Recall metric (0.972), not on the other aggregate metrics. This pattern challenges the "newer-is-always-better" approach that often serves as the default in architecture selection, underscoring the importance of empirical evaluation specific to the intended application context [34].

In the context of national air defense system development, these findings provide what was previously not explicitly available in the local literature: a model-selection framework that can be directly mapped to different deployment scenarios. YOLOv5s emerges as the most rational choice for edge-based ISR platforms such as reconnaissance drones or military-grade embedded systems, where memory limitations, power constraints, and inference speed are primary considerations [35]. YOLOv8s is better suited for ground stations with adequate computational resources that require balanced performance across all metrics. YOLOv11s, with its superior Recall, is most valuable in wide-area surveillance contexts or adversarial scenarios where missed detections carry serious tactical consequences.

Equally important insights emerge from the per-class analysis, which reveals patterns beyond initial predictions. Fighter aircraft such as F-22, F-16, and F-15 turn out to be the most difficult classes to detect, with mean AP ranging from 0.705 to 0.788, whereas large bomber and transport aircraft such as TU-95, B-52, and TU-22 achieve AP values above 0.85. This pattern indicates that inter-class visual similarity is a more dominant factor than sample imbalance fighter aircraft share similar planform characteristics in top-down imagery (wing geometry, fuselage ratios, twin-tail configurations) that complicate fine-grained discrimination. The practical implication is that future performance improvements should focus not only on class balancing but also on fine-grained classification techniques specific to visually similar aircraft categories.

On the efficiency dimension, the trade-off comparison between accuracy and parameter count yields directly

actionable insights. YOLOv5s, with 7.2M parameters, achieves accuracy equivalent to YOLOv8s, which has 11.2M parameters 56% larger. This demonstrates that increased architectural complexity does not always correlate linearly with performance gains, particularly on controlled datasets like MAR20. For developers of aircraft detection systems based on embedded platforms such as MALE (Medium-Altitude Long-Endurance) drones or optical systems at airbases, this finding provides a more realistic basis for computational resource planning: the lightest model is already sufficient to achieve competitive detection performance.

Several limitations of this study should be transparently acknowledged to guide further research. First, each architecture was trained and evaluated under a single training run with a fixed random seed; the reported performance differences therefore reflect a single realization and have not been tested for statistical significance through repeated runs (for example, five seeds with paired tests on per-image AP). Differences on the order of those observed between YOLOv5s, YOLOv8s, and YOLOv11s could in principle fall within run-to-run variance, and future replication with multiple seeds is needed before the ranking can be treated as robust. Second, computational performance was characterized only through parameter count; FLOPs, GPU/CPU inference latency, achievable frames-per-second, and memory footprint were not measured, which means the edge-deployment recommendations are inferential rather than directly validated on representative hardware such as NVIDIA Jetson modules or comparable embedded platforms. Third, error analysis at the per-class level identifies fighter aircraft (F-22, F-16, F-15) as the most challenging category but does not include qualitative inspection of failure modes for example, Grad-CAM or similar attribution maps on misclassified instances, or confusion analysis between specific fighter pairs so the visual-similarity hypothesis is supported by aggregate AP values rather than direct evidence of where each model attends. Fourth, the dataset itself (MAR20) covers 20 classes from remote sensing imagery and does not include camouflaged aircraft, partially obscured targets, oblique or in-flight imagery, or adversarial weather conditions, all of which are operationally relevant scenarios in real ISR settings. Fifth, only the small variant of each YOLO family was evaluated; the relative ordering may shift for m, l, and x variants with greater representational capacity. Each of these limitations defines a concrete direction for future work below.

VI. CONCLUSION

Based on the experimental results, this study concludes that all three evaluated YOLO architectures YOLOv5s, YOLOv8s, and YOLOv11s demonstrate strong capabilities for military aircraft detection on remote sensing imagery, consistently achieving mAP@0.5 above 0.988 on the MAR20 dataset. However, the most important finding lies not in this aggregate figure but in the distinct strengths each architecture exhibits: YOLOv5s excels in combining the highest accuracy with the best parameter efficiency (Precision 0.985, F1-Score 0.978,

with only 7.2M parameters), YOLOv8s offers the most balanced performance across metrics, while YOLOv11s delivers the highest Recall for scenarios requiring maximum detection coverage. This pattern suggests that the selection of detection architectures should not be based solely on version recency but tailored to the specific operational profile of the system being developed, while we note that broader confirmation would benefit from repeated runs with statistical testing and evaluation across additional datasets.

The study also reveals an unexpected pattern at the per-class level: fighter aircraft such as the F-22, F-16, and F-15 emerge as the most challenging categories to detect due to their high top-down visual similarity, whereas larger aircraft such as the TU-95 and B-52 can be detected with high accuracy thanks to their distinctive structural features. This insight provides a methodological contribution by highlighting that inter-class visual similarity is a more dominant factor than sample imbalance in determining detection difficulty for fine-grained military aircraft classification.

The contribution of this study extends beyond providing the previously unavailable comparative analysis of YOLOv5/v8/v11 in the literature; it also offers a practical model-selection framework for ISR system development in Indonesia. For deployment on edge platforms such as reconnaissance drones and military-grade embedded systems, YOLOv5s appears to be a strong candidate, while YOLOv8s and YOLOv11s may be better suited for ground station scenarios and wide-area surveillance, pending validation through on-device latency benchmarking. Furthermore, this study contributes to the broader literature on object detection in remote sensing imagery by emphasizing the importance of more granular evaluation approaches focusing not only on aggregate metrics but also on per-class characteristics and accuracy–efficiency trade-offs to produce more accurate and context-specific deployment recommendations. Future research is encouraged to extend this work along several concrete directions: (1) replicating the comparison with multiple random seeds (for example, five runs per architecture) and applying paired statistical tests on per-image AP to confirm whether the observed ranking between YOLOv5s, YOLOv8s, and YOLOv11s is significant or falls within training variance; (2) extending the benchmark to the m, l, and x variants of each architecture to characterize how the accuracy–capacity trade-off evolves with model scale; (3) profiling FLOPs, inference latency, achievable frames-per-second, and memory footprint on representative edge hardware such as NVIDIA Jetson devices to convert the present accuracy-based guidance into validated deployment recommendations; (4) adding qualitative error analysis through Grad-CAM or comparable attribution methods on misclassified fighter aircraft to verify whether the inter-class visual-similarity hypothesis is supported at the feature-attention level; and (5) evaluating cross-domain generalization on operational imagery with camouflaged, partially obscured, in-flight, and weather-degraded targets to assess robustness beyond the MAR20 distribution. Together, these directions would transform the present comparative study into readily

deployable implementation guidelines for Indonesia's national air defense systems.

REFERENCES

- [1] Z. F. Tiwery, I. H. Hanafi, and W. A. Riry, “Pelanggaran Wilayah Udara Indonesia Oleh Pesawat Asing Menurut Hukum Internasional,” *Balobe Law Journal*, vol. 4, no. 1, p. 35, Apr. 2024, doi: 10.47268/balobe.v4i1.2046.
- [2] K. M. Elgamily, M. A. Mohamed, A. M. Abou-Taleb, and M. M. Ata, “Enhanced object detection in remote sensing images by applying metaheuristic and hybrid metaheuristic optimizers to YOLOv7 and YOLOv8,” *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-89124-8.
- [3] S. Gui, S. Song, R. Qin, and Y. Tang, “Remote Sensing Object Detection in the Deep Learning Era A Review,” Jan. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/rs16020327.
- [4] J. Terven, D. M. Córdova-Esparza, and J. A. Romero-González, “A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS,” Dec. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/make5040083.
- [5] H. Juwiantho and J. E. Saputra, “Robust Balinese Script Detection for Transliteration Using YOLOv8,” *Teknika*, vol. 15, no. 1, pp. 178–185, Mar. 2026, doi: 10.34148/teknika.v15i1.1466.
- [6] M. Mao and M. Hong, “YOLO Object Detection for Real-Time Fabric Defect Inspection in the Textile Industry: A Review of YOLOv1 to YOLOv11,” Apr. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/s25072270.
- [7] G. Cinarer, Y. S. Taspinar, and M. Zeybek, “Detection of Aerial Vehicles Using Satellite Imagery: Comparative Analysis of U-Net Segmentation Model and YOLO Object Detection Model,” *International Journal of Aerospace Engineering*, vol. 2025, no. 1, 2025, doi: 10.1155/ijae/5599522.
- [8] R. H. Pratama, Ernawati, and W. O. KZ, “Modification of YOLOv11 Architecture with Convolutional Block Attention Module Integration for Batik Besurek Motif Detection,” *Teknika*, vol. 15, no. 1, pp. 103–112, Mar. 2026, doi: 10.34148/teknika.v15i1.1449.
- [9] X. V. Tandianto, D. H. Bangkalang, and N. Setiyawati, “Comparison and Implementation of CNN Facial Emotion Recognition Model with Hyperparameter Analysis on Multiple Datasets,” *Teknika*, vol. 15, no. 1, pp. 75–83, Mar. 2026, doi: 10.34148/teknika.v15i1.1445.
- [10] G. Yao, S. Zhu, L. Zhang, and M. Qi, “HP-YOLOv8: High-Precision Small Object Detection Algorithm for Remote Sensing Images,” *Sensors*, vol. 24, no. 15, Aug. 2024, doi: 10.3390/s24154858.
- [11] B. Chen, Z. Ma, and X. Li, “Boosting Detection Accuracy: An Enhanced YOLOv8 for Small Target



- Detection in Remote Sensing,” *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, Dec. 2025, doi: 10.1007/s44196-025-00895-4.
- [12] W. Yu *et al.*, “MAR20 : A benchmark for military aircraft recognition in remote sensing images,” *National Remote Sensing Bulletin*, vol. 27, no. 12, pp. 2688–2696, 2023, doi: 10.11834/jrs.20222139.
- [13] X. Cao, Y. Zhang, S. Lang, and Y. Gong, “Swin-Transformer-Based YOLOv5 for Small-Object Detection in Remote Sensing Images,” *Sensors*, vol. 23, no. 7, Apr. 2023, doi: 10.3390/s23073634.
- [14] S. Rattanaamporn, A. Perera, A. Nguyen, T. B. Ngo, and J. Chahl, “Drone Imaging and Sensors for Situational Awareness in Hazardous Environments: A Systematic Review,” Oct. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/jsan14050098.
- [15] C. Bai, X. Bai, and K. Wu, “A Review: Remote Sensing Image Object Detection Algorithm Based on Deep Learning,” Dec. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/electronics12244902.
- [16] B. Zhao, Y. Zhou, R. Song, L. Yu, X. Zhang, and J. Liu, “Modular YOLOv8 optimization for real-time UAV maritime rescue object detection,” *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-75807-1.
- [17] M. Hussain, “YOLOv1 to v8: Unveiling Each Variant-A Comprehensive Review of YOLO,” *IEEE Access*, vol. 12, pp. 42816–42833, 2024, doi: 10.1109/ACCESS.2024.3378568.
- [18] F. Chen, Y. Zhang, L. Fu, H. Xie, Q. Zhang, and S. Bi, “From YOLOv5 to YOLOv8: Structural Innovations and Performance Improvements,” *International Journal of Computer Science and Information Technology*, vol. 6, no. 1, pp. 93–100, May 2025, doi: 10.62051/ijcsit.v6n1.12.
- [19] Y. Zhao, F. Sun, and X. Wu, “FEB-YOLOv8: A multi-scale lightweight detection model for underwater object detection,” *PLoS One*, vol. 19, no. 9 September, Sep. 2024, doi: 10.1371/journal.pone.0311173.
- [20] P. Lu, X. Hao, W. Li, C. Yi, R. Kong, and T. Wang, “ECF-YOLO: An enhanced YOLOv8 algorithm for ship detection in SAR images,” *Electronic Research Archive*, vol. 33, no. 5, pp. 3394–3409, 2025, doi: 10.3934/era.2025150.
- [21] H. Huelss, “Transcending the fog of war? US military ‘AI’, vision, and the emergent post-scopie regime,” *European Journal of International Security*, vol. 10, no. 2, pp. 190–210, May 2025, doi: 10.1017/eis.2024.21.
- [22] K. Helkala, G. Østby, G. Reichberg, and M. Metcalf, “The PRC considers military AI ethics: Can autonomy be trusted?,” Oct. 2022. doi: 10.3389/fdata.2022.991392.
- [23] A. Downey, “The alibi of AI: algorithmic models of automated killing,” *Digital War*, vol. 6, no. 1, Dec. 2025, doi: 10.1057/s42984-025-00105-7.
- [24] X. Chen *et al.*, “DET-YOLO: An Innovative High-Performance Model for Detecting Military Aircraft in Remote Sensing Images,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 17753–17771, 2024, doi: 10.1109/JSTARS.2024.3462745.
- [25] S. El Ghazouali, A. Gucciardi, F. Venturini, N. Venturi, M. Rueeggsegger, and U. Michelucci, “FlightScope: An Experimental Comparative Review of Aircraft Detection Algorithms in Satellite Imagery,” Dec. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/rs16244715.
- [26] Z. Li *et al.*, “RSI-YOLO: Object Detection Method for Remote Sensing Images Based on Improved YOLO,” *Sensors*, vol. 23, no. 14, Jul. 2023, doi: 10.3390/s23146414.
- [27] J. Wang, Z. Bai, X. Zhang, and Y. Qiu, “A Lightweight Remote Sensing Aircraft Object Detection Network Based on Improved YOLOv5n,” *Remote Sens. (Basel)*, vol. 16, no. 5, Mar. 2024, doi: 10.3390/rs16050857.
- [28] T. Adli, D. Bujaković, B. Bondžulić, M. Laidouni, and M. Andrić, “Comparative analysis of YOLO algorithms for aircraft detection in remote sensing images,” Centre for Evaluation in Education and Science (CEON/CEES), Oct. 2024, pp. 326–331. doi: 10.5937/oteh24059a.
- [29] I. E. Tampu, A. Eklund, and N. Haj-Hosseini, “Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images,” *Sci. Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1038/s41597-022-01618-6.
- [30] J. Chen, S. Pan, Y. Chan, Y. Ni, and D. Ye, “A new method based on YOLOv5 and multiscale data augmentation for visual inspection in substation,” *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-60126-2.
- [31] K. Nakamura, B. Derbel, K. J. Won, and B. W. Hong, “Learning-rate annealing methods for deep neural networks,” *Electronics (Switzerland)*, vol. 10, no. 16, Aug. 2021, doi: 10.3390/electronics10162029.
- [32] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. Da Silva, “A comparative analysis of object detection metrics with a companion open-source toolkit,” *Electronics (Switzerland)*, vol. 10, no. 3, pp. 1–28, Feb. 2021, doi: 10.3390/electronics10030279.
- [33] R. Khanam, T. Asghar, and M. Hussain, “Comparative Performance Evaluation of YOLOv5, YOLOv8, and YOLOv11 for Solar Panel Defect Detection,” *Solar*, vol. 5, no. 1, Mar. 2025, doi: 10.3390/solar5010006.
- [34] R. Sapkota *et al.*, “YOLO advances to its genesis: a decadal and comprehensive review of the You Only Look Once (YOLO) series,” *Artif. Intell. Rev.*, vol. 58, no. 9, Sep. 2025, doi: 10.1007/s10462-025-11253-3.
- [35] N. Habash, A. A. Alqumsan, and T. Zhou, “Recent Real-Time Aerial Object Detection Approaches, Performance, Optimization, and Efficient Design Trends for Onboard Performance: A Survey,” Dec. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/s25247563.