International Journal of Combinatorial Optimization Problems and Informatics, 17(2), 2026, 148-162. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i2.1269

www.editada.org

A Comparative Study of Fusion Methods for Firearm Detection

Kevin Jorge Montes Lorenzo, Francisco José Moo Mena, Antonio Armando Aguileta Güemez
Universidad Autónoma de Yucatán
a22216420@alumnos.uady.mx, mmena@correo.uady.mx, aaguilet@correo.uady.mx

Abstract. In a previous study, we compared the performance of two well-known models, YOLOv8 and RT-DETR, for firearm detection. The results showed that YOLOv8 achieved superior performance, while RT-DETR also produced significant results. These findings suggested the potential to further improve detection by exploring result fusion methods. Unlike the original comparative analysis, this work investigates how integrating the outputs of multiple detectors can enhance both accuracy and robustness in firearm identification. The existing literature on result fusion, as explicitly applied to this field is scarce, leaving a promising line of research open. In this context, strategies such as averaging results, selecting the best detector, and time-weighted as well as real-time weighted methods are analyzed. The objective is to demonstrate that result fusion represents an effective method for enhancing the performance of object detection systems in complex scenarios.

Keywords: firearm detection, YOLOv8, RT-DETR, result fusion, object detection, robustness.

Article Info
Received Dec 26, 2025
Accepted Janc 11, 2026

1 Introduction

Each year, the number of crimes involving firearms has gradually increased. One way to minimize such attacks and pursue and arrest these aggressors is through the use of surveillance systems, which can provide the necessary evidence for this purpose. Closed-circuit recording systems are used as security tools, but their response is often slow and prone to errors because humans monitor them.

In recent years, advances in the field of deep learning and artificial intelligence have enabled the development of highly accurate and efficient object detection systems in images. These systems are based on algorithms that automatically learn to identify relevant patterns in images that indicate the presence of these objects. The advantage of these detection systems is that they are capable of detecting such objects in a wide variety of situations and environments, including low-quality images, poor lighting conditions, or with foreground objects that may obstruct the view of the image. Moreover, these systems can easily adapt to new situations and contexts, making them ideal for use in dynamic and ever-changing environments.

The detection of firearms in images is a critical task in public safety and law enforcement, as the presence of weapons in certain situations can pose a threat to people's safety. In summary, the development of a machine learning or artificial intelligence-based model for detecting weapons in images is of great importance for both public and private security. These systems allow for faster and more accurate detection of weapons in images, which can help prevent dangerous and potentially deadly situations. Firearms were chosen for detection because they are the most commonly used and represent a greater threat due to their ease of use in mass attacks.

In this context, it is crucial to compare different state-of-the-art object detection models, such as YOLOv8 (VK_Venkatkumar, 2023) and RT-DETR (Ultralytics, 2024), to determine which one offers better results in the specific task of detecting firearms in images. Both models represent recent advances in the field of artificial intelligence, but they differ in their architectures, efficiency, and accuracy. Conducting this comparison is especially important because, to date, previous works have not directly compared these models for this specific task. This research will help identify the model that offers the best performance, thereby improving the effectiveness of surveillance systems and ultimately increasing public safety.

Additionally, this work also explores result fusion methods as a complementary strategy to enhance performance. The idea behind result fusion is that, by combining the outputs of multiple models or detection approaches, it is possible to compensate for the weaknesses of each method while reinforcing their strengths. This strategy has been successfully applied in other areas of computer vision, where the fusion of multiple classifiers or detectors has led to more robust and accurate systems. In the context of firearm detection, the objective is to determine whether the fusion of results between different models can further increase detection accuracy, reduce false positives, and provide more reliable outcomes in diverse and challenging scenarios.

The motivation behind this research stems from the need to continually improve the technological tools used to prevent and mitigate the impact of armed violence in society.

A relevant aspect is that, despite the growing interest in firearm detection systems through computer vision, there are few studies that specifically analyze the use of result fusion methods in this domain. This scarcity presents a significant research opportunity, as integrating predictions from different models can be crucial to reducing uncertainties and leveraging the unique strengths of each detector.

Among decision fusion strategies, we consider approaches such as the average of results, the selection of the best detector, as well as time-weighted and real-time weighted methods, which enable the integration of multiple model outputs more stably and reliably. Applying these techniques in the context of firearm detection aims not only to increase accuracy, but also to improve the robustness of the system when facing complex and diverse scenarios.

In this sense, the present work extends beyond model comparison. It introduces a new perspective focused on evaluating how result fusion can enhance performance in firearm detection, representing a novel and complementary contribution to the previous study.

In the following sections, the structure of this work is presented to provide a clear roadmap for the reader. Section 2, Concepts and Previous Work, introduces key concepts related to fusion methods in artificial intelligence and reviews previous studies that have addressed similar topics, highlighting their approaches and findings. Section 3, Methodology, details the phases of the proposed methodology, including dataset preparation, preprocessing, model training, and the application of different fusion strategies. Section 4, Experimentation, describes the implementation of the methodology, explaining how experiments were conducted and the configurations used. Finally, Section 5, Results, presents the outcomes of the experiments, analyzes the performance of the fusion methods, and discusses their implications for practical firearm detection scenarios. Finally, Section 6, Conclusion, summarizes the main findings of this research, and outlines its contributions.

2 Concepts and Previous Work

In artificial intelligence, information fusion (Pereira, Salazar, & Vergara, 2024) refers to the systematic process of combining data from multiple sources or models to produce a superior output—typically more accurate, robust, or reliable than that derived from any individual source. This paradigm has delivered notable benefits across various domains, including sensor integration for autonomous driving, multimodal cognition, medical diagnostics, and object detection. Fusion methods in artificial intelligence can be classified into different levels depending on the stage at which data or predictions are integrated:

- Data-level fusion: directly combines raw information from multiple sources.
- Feature-level fusion: merges intermediate representations extracted by different models into a joint vector.
- Result-level fusion: integrates the final predictions of several detectors using rules such as averaging, majority voting, or maximum confidence.
- Decision-level fusion: considers each output as an independent decision and then combines them into a global decision, using schemes such as best-choice selection, weighted voting, or dynamic strategies.

In this article, the focus is on result-level fusion, employing methods such as result averaging, best detector selection, and time-weighted and real-time weighted variants, to enhance firearm detection.

An Analysis on Ensemble Learning Optimized Medical Image Classification with Deep Convolutional Neural Networks (Müller et al., 2022): This work analyzes the use of ensemble techniques applied to medical image classification, a domain where accuracy is critical to support clinical diagnoses. The authors compare different model combination approaches, including bagging, stacking, and aggregation-based methods such as simple and weighted averaging. The results show that stacking stands out as the most effective strategy, as it enables a more sophisticated integration of information from multiple classifiers, leading to significant accuracy improvements. However, the study also highlights that simple statistical methods, such as averaging outputs, provide highly competitive performance with lower computational cost, making them attractive for practical implementations in hospital settings. This research demonstrates the utility of ensemble learning in strengthening classification in medical contexts, where prediction errors must be minimized as much as possible.

A Hybrid Deep Learning Framework with Decision-Level Fusion for Breast Cancer Survival Prediction (Othman et al., 2023): This article proposes a hybrid deep learning framework for predicting survival in breast cancer patients, in which decision-level fusion techniques are employed. The approach combines multiple independent classifiers, whose outputs are integrated through aggregation methods to generate a more reliable final prediction. The research emphasizes that fusion at this level allows leveraging the strengths of different models while reducing their individual weaknesses, which is especially relevant in clinical contexts where highly reliable estimations are required. Experimental results indicate significant improvements compared to particular models, underlining the importance of result fusion as a key resource in developing medical decision support and diagnostic systems in oncology.

Performance Evaluation of Different Decision Fusion Approaches for Image Classification (Alwakeel et al., 2023): This study compares various decision fusion methods applied to image classification tasks, aiming to assess how model combination improves performance compared to the use of individual models. Among the techniques analyzed are logit summation, majority voting, max/min pooling of outputs, and other aggregation variants. The experiments demonstrate that not all fusion methods offer the same benefits: for instance, majority voting tends to smooth out errors but does not always accurately capture model confidence, whereas logit summation leverages probabilistic information more effectively, resulting in superior performance in most evaluated cases. Furthermore, the study highlights that the effectiveness of each method may vary depending on the dataset and base architecture, reinforcing the importance of selecting the appropriate fusion technique according to the application context.

Decision-level fusion detection method of visible and infrared images under low light conditions (Hu, Jing, & Wu, 2023): This work addresses the challenge of object detection under low-light conditions, where visible images alone have limited performance. To overcome this difficulty, the authors propose a decision-level fusion method that combines visible and infrared images, leveraging their complementary characteristics. The detection architecture used as the backbone is YOLOX, which enables fast and accurate detections in both image modalities.

The article (Ouyang H, 2024) presents DEYO. This real-time object detection model innovatively integrates the DETR (Detection Transformer) and YOLO (You Only Look Once) architectures. This integration is not limited to a simple combination of two models. Still, it is carried out through a step-by-step training approach, designed to leverage the strengths of both methods: DETR’s ability for global and contextual detection, combined with YOLO’s efficiency and speed. This gradual procedure enables DEYO to overcome several limitations of traditional DETR models, which often require extensive pretraining on large datasets, such as ImageNet, and exhibit instability during the initial stages of training, particularly when handling complex object assignments. Regarding fusion methods, DEYO employs a type of architectural and feature fusion, where the detector’s backbone and neck are initialized with a pretrained detector and then combined with a transformer-style decoder trained independently in a second stage. This strategy can be considered a form of feature and model fusion, as it integrates the information processed by YOLO into the DETR framework, achieving more stable, accurate, and efficient real-time detection without the need for additional datasets or excessive computational resources.

3 Methodology

In Fig. 1, the phases that comprise the methodology developed for analyzing the fusion methods of interest are presented. The objective of each phase is described next. In the experimentation section, the implementation of this methodology for this study is explained.

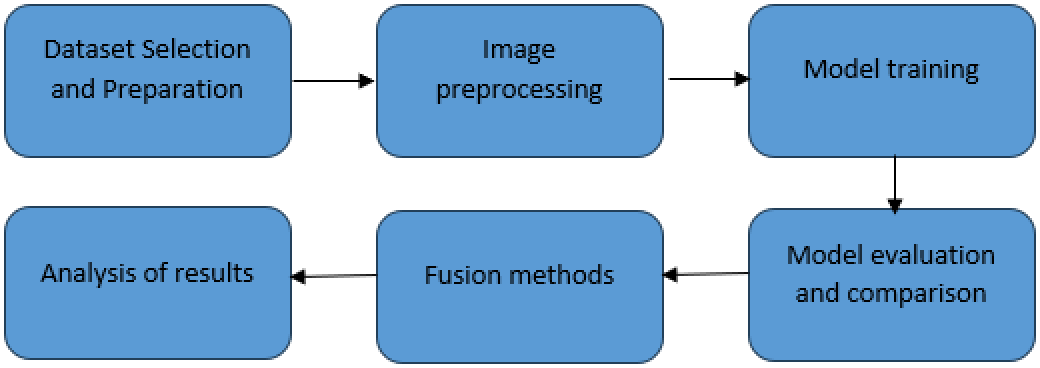

Fig. 1. Methodology sequence.

Dataset Selection and Preparation:

- A comprehensive search was carried out to identify publicly available datasets that contain images of firearms under different conditions (lighting, orientation, occlusion, and background complexity). The goal of this phase was not only to gather a sufficient number of images but also to ensure diversity and representativeness, which are essential for training robust detection systems. To achieve this, datasets from various sources were unified into a single collection, with careful review of the annotation formats to ensure consistency and compatibility across all samples. This step ensures that subsequent training phases are based on a high-quality and standardized dataset.

Image preprocessing:

- Once the dataset was consolidated, preprocessing was applied to ensure uniformity and enhance the quality of the input images. In cases where the dataset size was limited, data augmentation techniques such as rotation, flipping, scaling, and brightness adjustments were employed. These methods increase dataset variability, reduce overfitting, and help the models generalize better when encountering unseen firearm images in real-world scenarios.

Model training:

- Two state-of-the-art models, YOLOv8 and RT-DETR, were selected for training due to their proven accuracy and efficiency in object detection tasks. During this phase, both models were trained on the prepared dataset using optimal hyperparameters determined through iterative experimentation. The training process was carefully monitored with validation checkpoints to prevent overfitting and to assess learning progress. The inclusion of two different architectures allows for later exploration of their complementary strengths, which is essential for the subsequent application of fusion strategies.

Model evaluation and comparison:

- After training, both models were rigorously evaluated using a validation set to quantify their performance. Evaluation metrics included accuracy, Intersection over Union (IoU), precision, recall, and inference time per image. These metrics provide a comprehensive understanding of not only how well each model detects firearms but also how efficiently they can be deployed in real-time applications. A comparative analysis was then performed between YOLOv8 and RT-DETR, highlighting their relative strengths and weaknesses in terms of detection speed, robustness against occlusions, and accuracy in complex backgrounds.

Fusion methods:

- Once the models have produced their predictions, different fusion strategies will be applied to combine their outputs.
- The fusion methods include simple averaging, weighted average over time, real-time weighted average, and selection of the best result per image.
- These strategies are designed to exploit the strengths of both models, aiming to improve consistency and robustness in firearm detection.

Analysis of results:

- The results will be analyzed to determine the effectiveness of each model in the specific task of firearm detection.
- Conclusions will be drawn about the applicability of each model in realistic security environments.

4 Experimentation

Dataset Selection and Preparation:

A thorough search was conducted for public datasets containing firearm images (DASCI, n.d.; Roboflow Universe, n.d.). The selected datasets were combined to form a consolidated dataset. The final dataset consists of 3,274 images, divided into two subsets: 2,382 images (72.8%) for training and 892 images (27.2%) for validation. This split ensures fair evaluation and prevents overfitting because it ensures that the model is trained and fine-tuned on a subset of the data (the training set). At the same time, its performance is evaluated on a separate set (the validation set) that was not used during training. This helps obtain a more representative performance metric of how the model will generalize to unseen data. A test set was not used due to data scarcity. If the available data is limited, splitting it into three sets may reduce the training set size too much. In such cases, validation is prioritized to fine-tune the model, leaving testing as a future task on an unseen dataset. In some studies, the authors train and validate their models using their own data but evaluate the final performance using external test datasets. Fig. 2 shows some images from a dataset.

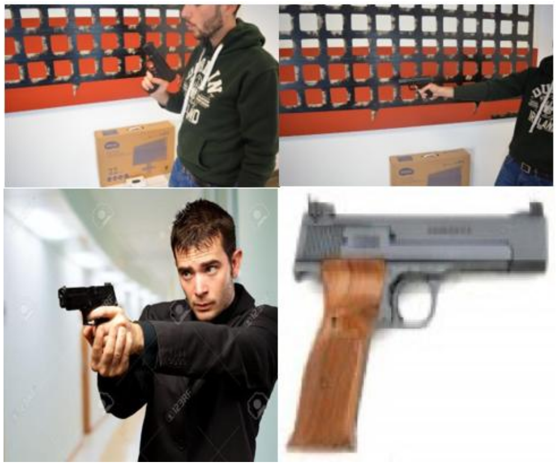

Fig. 2. Example image of the dataset.

Image preprocessing:

Each image underwent preprocessing, including image enhancement using pipelines and OpenCV, as many images were blurry or unclear (see Fig.3). Firearms were annotated using a standard annotation format (YOLO format) to ensure compatibility across models. The images in one dataset have dimensions of 416x416, while those in the other dataset are 1620x1080. The preprocessing performed included increasing brightness and sharpening all images, as some exhibited a certain level of blur. Another process applied was data augmentation, which both models use by default through Albumentations, a Python library designed for applying transformations and augmentations to images, particularly in deep learning tasks such as object detection, segmentation, and classification. This library enables operations such as rotations, cropping, brightness and contrast adjustments, blurring, and more, to enhance the robustness of the models.

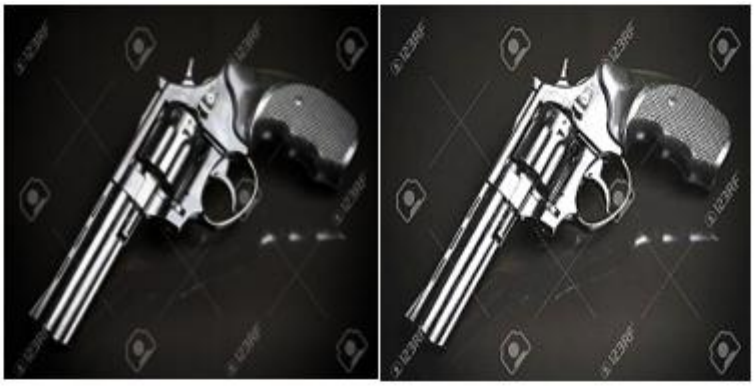

Fig. 3. Example image of the dataset after increasing the sharpness and clarity.

Additionally, data augmentation techniques, such as rotations, brightness changes, and cropping, were applied to enrich the dataset and improve model robustness against image variations (see Fig.4). The models automatically provided data augmentation, and the user could choose to implement it.

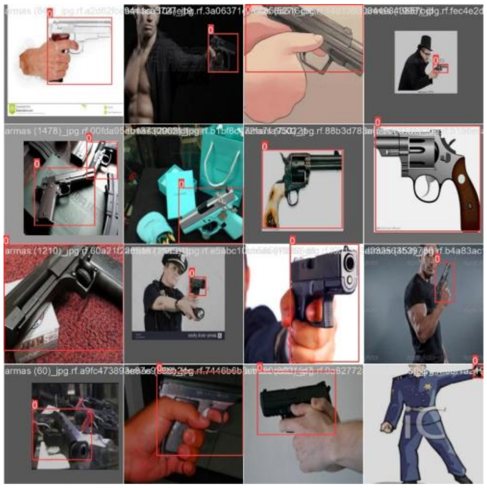

Fig. 4. Example image of the dataset with albumentations process.

Model training:

- Two state-of-the-art object detection models, YOLOv8 and RT-DETR, were selected and independently trained using the training set.
- An iterative approach was used during training, adjusting hyperparameters such as learning rate, number of epochs, and batch size to optimize model performance.
- Loss and accuracy metrics were monitored during the training process to ensure convergence and avoid overfitting.

Training Epoch Selection

Both models were tested using the same number of training epochs. The best results were obtained with:

- 100 epochs for YOLOv8
- 20 epochs for RT-DETR

These values were chosen as the final configurations for presentation in the study.

Model evaluation and comparison:

- After training, the models were evaluated using the validation set. Key metrics considered were accuracy, Intersection over Union (IoU), and inference time per image.
- A detailed comparison between YOLOv8 and RT-DETR was conducted, focusing on detection speed and accuracy in identifying firearms. This analysis identified the strengths and weaknesses of each model in practical scenarios.

Fusion methods:

After obtaining the individual predictions from YOLOv8 and RT-DETR, several fusion strategies were applied to combine their outputs.

The methods used included simple averaging, time-weighted averaging, real-time weighted averaging, and best-result selection.

These fusion approaches aimed to exploit the complementary strengths of both models, improving the robustness of detection and potentially achieving higher IoU scores compared to individual models.

Analysis of results:

- The results were thoroughly analyzed to determine the effectiveness of each model in the specific task of firearm detection. Both qualitative and quantitative analyses were conducted, including visual examples of successful and failed detections.
- Conclusions were drawn about the applicability of each model in realistic security environments, considering factors such as accuracy, speed, and implementation complexity.

To ensure a fair comparison between models or configurations, it is crucial to control variables that can affect performance. In this case, hyperparameters such as learning rate, batch size, and optimizer were kept constant, ensuring that any difference in performance is due exclusively to the number of epochs and not to other factors. If other hyperparameters, such as learning rate or momentum, had been modified, their impact would have mixed, making it harder to interpret the results. In many research studies, only one factor is varied at a time to assess its impact in isolation. In this case, only the number of epochs was modified to observe its effect on model convergence and avoid biases caused by other changes.

5 Results

This section presents the results obtained from training and evaluating the YOLOv8 and RT-DETR models in the task of detecting firearms in images. The results include key metrics such as mean Average Precision (mAP), Precision, Recall, and the evolution of losses during training and validation for both models.

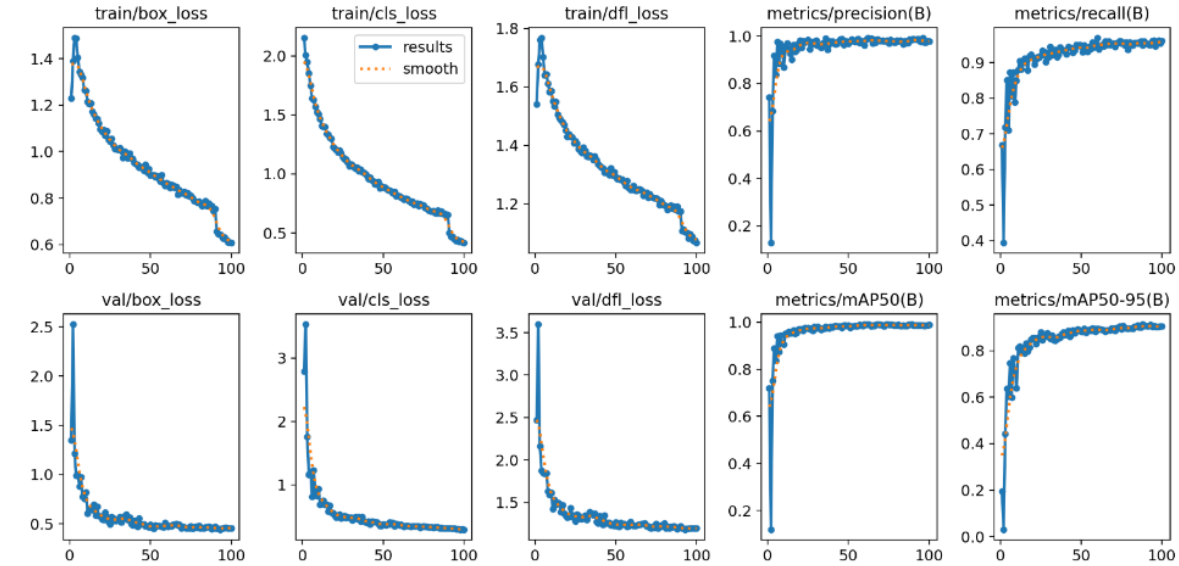

Fig. 5. Results obtained during YOLOv8 training for 100 learning epochs.

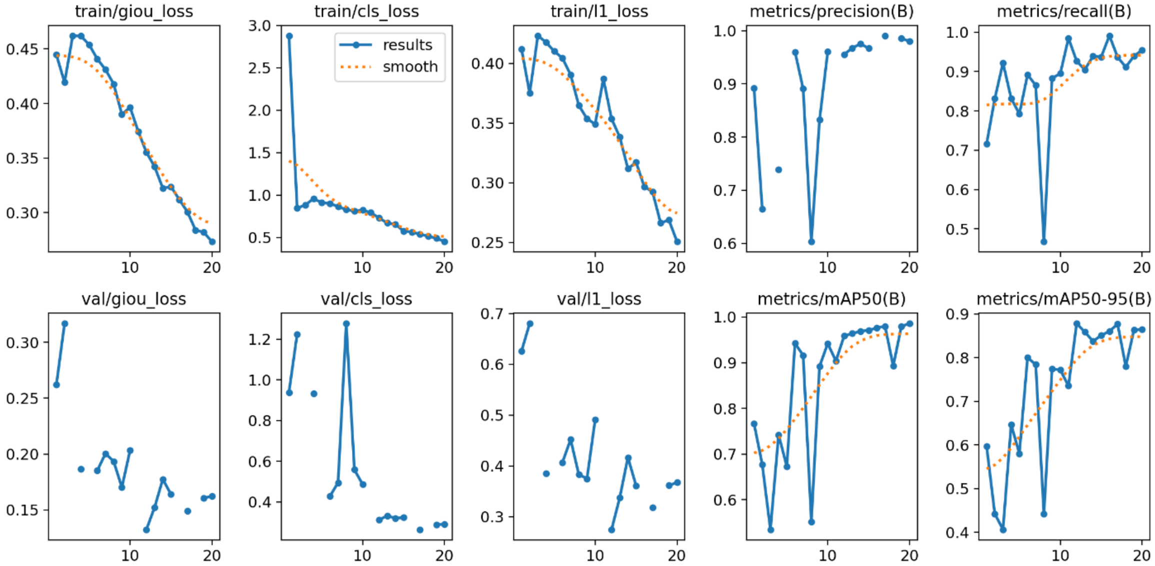

Fig. 6. Results obtained during RT-DETR training during 20 learning epochs.

These graphs (See Fig. 5, Fig. 6) indicate that the model effectively learned during both training and validation, with consistent improvements in key metrics such as precision, recall, and mAP, along with a decrease in associated losses.

The data augmentation strategy may cause isolated points, as some transformations can make specific images more challenging for the model, leading to spikes in validation loss.
Additionally, instability during training, possibly due to the learning rate or other hyperparameters, may contribute to fluctuations, especially in the early stages of training. These factors will be taken into consideration in future research.

metrics/precision: The model's precision remains high and relatively stable during training, indicating that it accurately predicts the correct classes.

metrics/recall: The recall, or sensitivity, improves rapidly in the early epochs and then stabilizes, showing good performance in recovering true positive objects.

metrics/mAP50: The "mean Average Precision" (mAP) with an IoU threshold of 0.50. This value increases quickly and then stabilizes at a high level, indicating that the model is doing a good job predicting the correct bounding boxes and classes.

5.1 Comparison

Precision and Recall: Both models exhibit competitive performance in terms of precision and recall, with RT-DETR showing a slight increase in precision at the end of training. However, YOLOv8 demonstrates superior stability throughout the training process (see Table 1).

Table 1. Comparison of Precision and Recall between RT-DETR and YOLOv8

Model	Precision	Recall
YOLOv8	0.99	0.97
RT-DETR	0.97	0.95

The main reason for the performance gap lies in the architectural differences between both models:

RT-DETR uses a transformer-based encoder-decoder architecture with an end-to-end detection mechanism.

- This allows it to learn contextual relationships between objects, which can enhance robustness.
- However, it requires more training iterations to stabilize because the attention mechanism is sensitive to learning rate and data distribution.
- The observed fluctuations in the precision and recall curves reflect this slower convergence behavior.

YOLOv8, on the other hand, employs a purely convolutional structure with decoupled head layers optimized for real-time detection.

- Its anchor-free prediction and path aggregation network (PAN) components improve localization and classification simultaneously.
- This leads to faster convergence, more stable precision and recall curves, and slightly higher performance metrics overall.

While RT-DETR offers a deeper understanding of object relationships and has strong generalization potential, YOLOv8 achieves higher and more stable precision–recall values due to its efficient convolutional design and extensive optimization for real-time applications.

mAP: Although both models achieve similar mAP50 values during validation (~0.86), RT-DETR shows more variability in metrics during the training process, while YOLOv8 maintains a more stable and consistent performance (see Table 2).

Table 2. Evaluation of Detection Accuracy (mAP) for RT-DETR and YOLOv8 Models

Model	mAP50	mAP50-95
YOLOv8	1.00	0.89
RT-DETR	0.98	0.86

The superior performance of YOLOv8, particularly in terms of stability and faster convergence, is mainly due to its anchor-based and convolution-optimized detection architecture, whereas RT-DETR implements a transformer-based anchor-free detection approach with attention-based decoding.

- Key factors explaining the difference: Detection mechanism. YOLOv8 uses an anchor-free approach that allows it to localize objects directly and efficiently by predicting bounding boxes without relying on predefined anchor points. RT-DETR, on the other hand, employs an anchor-free strategy that relies entirely on the transformer's attention mechanism to learn spatial relationships from scratch, which generally requires more data and longer training time.
- Convergence speed and stability: YOLOv8 exhibits a smooth and stable learning curve, quickly reaching mAP50 values close to 1.0. RT-DETR shows more noticeable oscillations, typical of attention-based models, as learning visual correspondences between regions is inherently more complex.
- Generalization capability: While RT-DETR achieves competitive mAP50-95 values, its overall performance is slightly lower due to higher sensitivity to variations in object scale and the detection of small objects.
- YOLOv8, through its Feature Pyramid Network (FPN) and Path Aggregation Network (PAN), enhances multi-scale detection, leading to better overall performance. The higher and more stable performance of YOLOv8 compared to RT-DETR is primarily attributed to its optimized convolutional and anchor-free architecture, which enables faster and more robust convergence.Meanwhile, RT-DETR, by relying on a fully end-to-end transformer-based design, offers greater theoretical contextual learning capacity, but at the cost of increased computational complexity and reduced training stability.
- Convergence: RT-DETR appears to converge faster in terms of precision and mAP, but its performance is less stable compared to YOLOv8, which shows a more gradual improvement and greater consistency.
- Losses: The losses in both models consistently decrease during training and validation. However, YOLOv8 presents higher losses (box, cls, and dfl) compared to RT-DETR, which may indicate greater complexity in optimizing this model.

In summary, YOLOv8 stands out for its stability and consistency during training and validation, making it a robust option for firearm detection. On the other hand, RT-DETR, while showing competitive performance in precision and mAP, has greater variability and seems to have less predictable behavior.

Example of Resulting Images:

Fig. 7 shows examples of images obtained with YOLOv8 and RT-DETR.

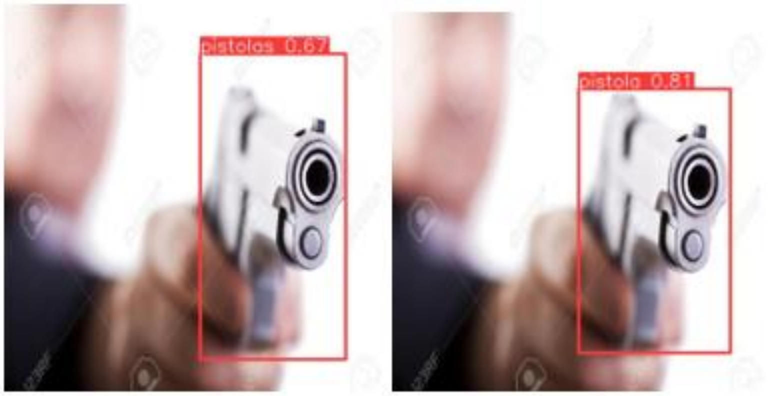

Fig. 7. Example of images obtained with YOLOv8 and RT-DETR, respectively.

5.2 Fusion of Results between YOLOv8 and RT-DETR

The objective of this stage was to evaluate whether combining the predictions of the YOLOv8 and RT-DETR models could provide improvements in terms of accuracy, robustness, and suitability in practical scenarios. The various fusion strategies implemented, along with their justifications, are described below.

- Direct comparison and best model selection. The first strategy involved a direct comparison between YOLOv8 and RT-DETR, considering key metrics such as precision, recall, mean Average Precision (mAP), and inference time. This comparison was performed on the same dataset under controlled conditions. The model with the best overall performance was selected as a baseline for future reference. Although simple, this strategy helps establish a clear and direct point of comparison. Advantage: enables a clear decision based on objective evidence. Limitation: completely discards useful information from the second model.
- Simple averaging of results. The second technique evaluated was the simple averaging of predictions from both models. Detection outputs (bounding boxes, classes, and confidence scores) were combined when they spatially overlapped (e.g., using IoU > 0.5). Averaging was applied to both the confidence scores and the bounding box positions. This strategy aims to mitigate individual errors of each model, assuming that their strengths complement each other. For example, if one model detects small firearms more accurately and the other performs better with larger ones, averaging could smooth these differences. Advantage: improves robustness against isolated false positives and false negatives. Limitation: may degrade performance if both models fail in different ways.
- Weighted averaging based on final performance. In this strategy, a weighted average was implemented, where each model contributed in proportion to its previously measured overall performance. For instance, if YOLOv8 achieved a mAP of 0.75 and RT-DETR a mAP of 0.65, YOLOv8 detections received greater weight in the combination. This technique seeks to capitalize on the strengths of the dominant model without completely discarding the useful detections of the second one. The weighting was applied to both confidence scores and bounding box positions. Advantage: balances global accuracy and model diversity. Limitation: Weighting is static, based on previous evaluations, rather than real-time performance.
- Real-time weighted averaging. The fourth strategy introduced a dynamic variant of weighted averaging, adjusting the weights in real-time based on the recent behavior of each model. Performance during inference was analyzed considering factors such as average confidence in the last N detections, prediction stability, and the number of matches with ground truth annotations (when available). This technique is beneficial in live surveillance systems, where environmental conditions can change rapidly. For example, if RT-DETR performs better in low-light night scenarios, the system can adjust its weighting to favor it in that temporary context. Advantage: adaptable to changing environmental conditions. Limitation: requires greater processing power and continuous monitoring of metrics.
- Best model selection. This strategy consists of selecting, for each prediction instance, the result provided by the model that demonstrates superior performance among the candidates—in this case, YOLOv8 and RT-DETR. Rather than averaging or combining the outputs of both models, this method relies on a comparative evaluation, whereby the most reliable detection is chosen dynamically from the two alternatives. This approach allows BMSelection to consistently capitalize on the strengths of each model, ultimately leading to an overall performance that surpasses either YOLOv8 or RT-DETR when considered independently.

For the evaluation of the methods proposed in this research, the same dataset specialized in firearm detection in images was used. This dataset includes a wide variety of scenarios that simulate real surveillance conditions, such as:

- Presence of firearms in indoor and outdoor environments
- Scenarios with one or multiple people
- Variations in lighting
- Different weapon positions

Each image in the dataset contains the necessary annotations in bounding box format, allowing for precise evaluation using metrics such as precision, recall, and mAP (mean Average Precision).

Although all evaluated methods used precisely the same set of images, the difference lies in how the results generated by each approach were processed and compared, depending on the methodological focus applied.

Results

In Fig. 8, the graph of averages with respect to the method used is observed, and it can be seen that they are pretty similar, indicating that the choice of the best method yields a better result.

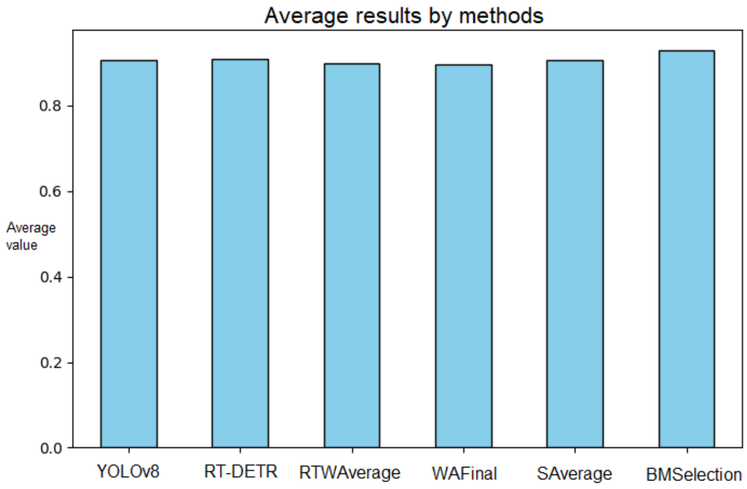

Fig. 8. Graph of averages.

Fig. 9 shows a comparative line graph of IoU per image, it can be observed that the methods YOLOv8, RT-DETR, RTWAverage, WAFinal, SAverage, and BMSelection show values that are pretty close to each other. This suggests that selecting the best method may lead to a slight improvement in overall results, although the differences are not significantly pronounced.

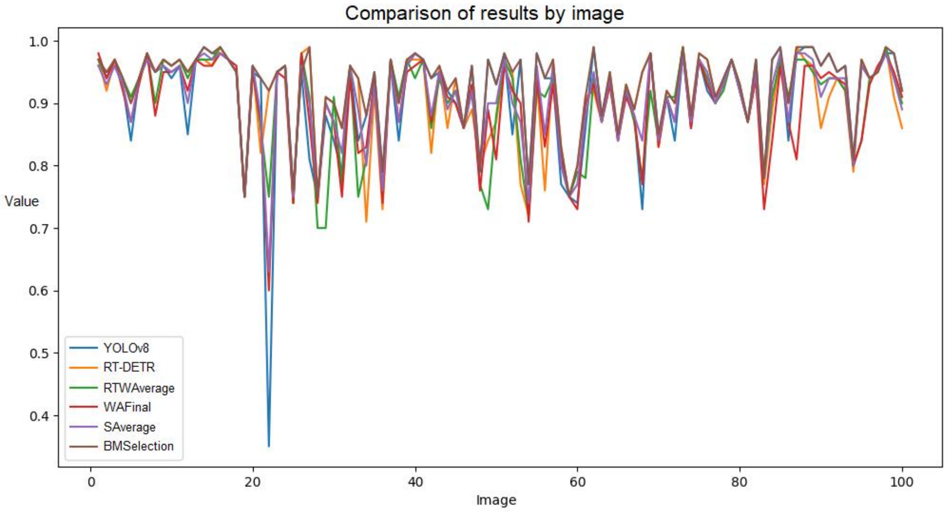

Fig. 9. Comparative line graph of IoU per image.

It is a comparative line chart of IoU results per image, showing how all the evaluated methods behave across the 100 images. All methods achieve high IoUs (0.85–0.98) in most images. This means the models are comparable in overall performance. There is no clear “loser” on average.

There are specific drops in certain images (e.g., between 20–30, 50–60, and 70–80).

Here we can see differences between models:

- YOLOv8 shows a sharp drop around image ~22.
- RTWAverage and WAFinal show notable drops at other points.
- RT-DETR seems more stable, although it also fluctuates.

This indicates that the models do not fail on the same images—each has strengths and weaknesses.

The “SAverage” and “BMSelection” lines are more stable:

- They stay higher and show fewer peaks.
- This makes sense because they represent an “average” or “best case,” which smooths out individual variations.

Direct comparison between YOLOv8 and RT-DETR:

- Both follow very similar curves.
- But there are images where one is clearly better than the other (point advantage).

Fig. 10 shows a correlation heatmap between detection methods.

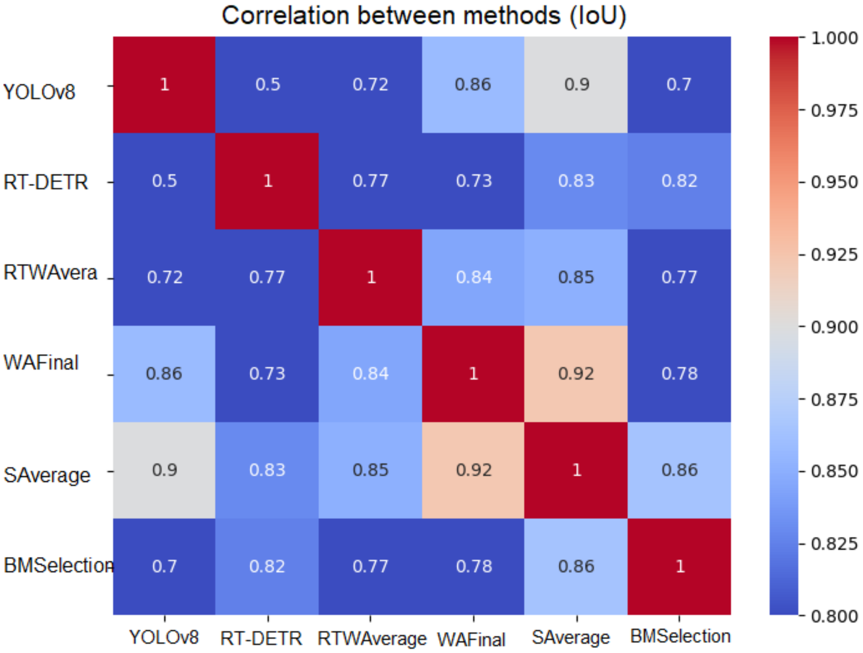

Fig. 10. Correlation heatmap between detection methods.

Correlation heatmap between different detection/evaluation methods (measured with IoU, Intersection over Union). Each cell represents the correlation between two different methods. Values range from -1 to 1:

- 1 = perfect correlation (both methods vary in the same way).
- 0 = no linear relationship.
- -1 = perfect inverse correlation.

The color scale goes from blue (low correlation ~0.8, in your case the minimum) to red (high correlation close to 1). On the diagonal (from top left to bottom right), the value is always 1, because each method is perfectly correlated with itself. YOLOv8 and SAverage (0.90) → highly correlated, meaning that when YOLOv8 has a good IoU, SAverage does too. YOLOv8 and RT-DETR (0.50) → low correlation, meaning that sometimes one works well while the other does not; their results do not always align.

WAFinal and SAverage (0.92) → highest correlation, suggesting these two methods measure very similar things. BMSelection and RT-DETR (0.82) → good relationship, though not the highest. Overall, most correlations are between 0.7 and 0.9, indicating that the methods tend to agree, except for some combinations, such as YOLOv8–RT-DETR, which show notable differences.

5.3 Statistical Evaluation: Friedman and Holm Tests

To ensure that the differences between the models and the fusion techniques were not due to chance, rigorous statistical tests were conducted, specifically the Friedman test followed by the Holm correction (Pereira et al., 2015). These tests allow for validating whether the improvements obtained in the experiments are statistically significant.

Friedman Test

The Friedman test is a non-parametric test used to compare three or more methods across multiple datasets. It is based on ranking the models instead of their absolute values, which makes it ideal in contexts where data normality cannot be assumed. In this work, the Friedman test was used to compare the performance of the following configurations:

- YOLOv8
- RT-DETR
- Simple average
- Weighted averaging based on final performance
- Real-time weighted averaging
- Best model selection

Each configuration was evaluated on multiple images and test scenarios, generating a set of rankings per trial. The null hypothesis of the test states that all methods have the same average performance. Rejecting this hypothesis confirms that there are significant differences among at least some of the evaluated methods.

Holm Correction

Once the Friedman test detects significant differences; it is necessary to conduct multiple pairwise comparisons to identify which methods differ significantly from one another. However, performing multiple comparisons increases the risk of committing type I errors (false positives).

To control this risk, the Holm correction was applied, an adjusted method that allows multiple comparisons without compromising statistical validity. Holm is an improvement over the Bonferroni method and works by sequentially adjusting the p-values of pairwise comparisons.

In this research, fusion techniques were specifically compared against individual models. The Holm correction provided statistical evidence that fusion strategies based on selecting the best result offered significant improvements over individual models in terms of mAP and correct detection rate.

Relevance of Statistical Tests

The combined use of Friedman and Holm strengthens the analysis since

•Friedman evaluates overall differences between methods without assuming normal distribution.

• Holm enables controlled pairwise comparisons, reducing the risk of misinterpretations due to multiple testing.

Thanks to these tests, it was possible to strongly support that fusion by selecting the best not only showed empirical improvements but also statistically significant differences, justifying its adoption as the preferred approach for dynamic surveillance scenarios involving firearm detection.

Fig. 11 shows results obtained by applying the Friedman and Holm tests. Each box represents the variability of the ranking assigned to each method, while the red values indicate the mean ranking.

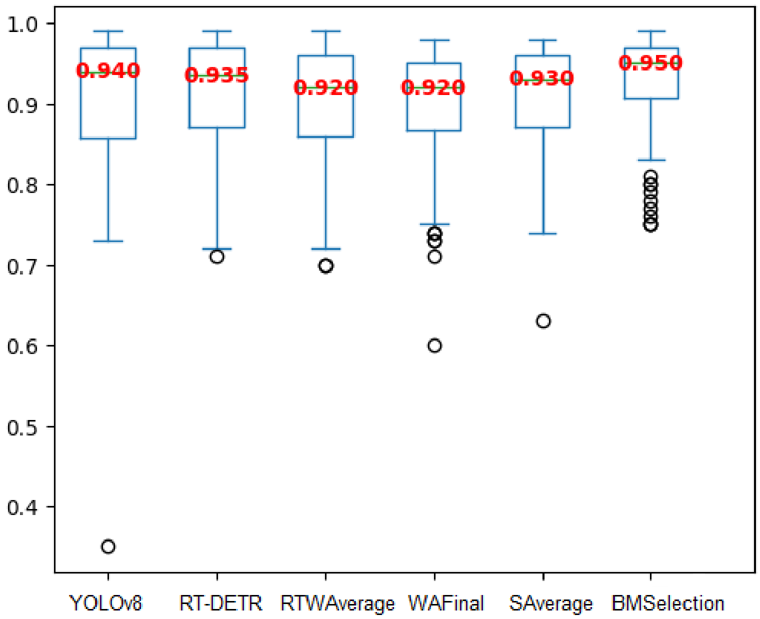

Fig. 11. Results obtained by applying the Friedman and Holm tests.

In the figure Fig.11, the methods YOLOv8, RT-DETR and fusion techniques (RTWAverage, WAFinal, SAverage, BMSelection) are compared. The metrics shown correspond to average performance values along with their dispersion.
YOLOv8 (0.940) and RT-DETR (0.935) are the individual baseline models. As observed, YOLOv8 achieves superior performance on its own. However, the fusion techniques also obtain competitive results: for instance, BMSelection reaches the highest value (0.950), even surpassing YOLOv8. Other fusions (RTWAverage and WAFinal with 0.920, SAverage with 0.930) exhibit intermediate performances, suggesting that not all fusion strategies are equally effective.

In general, the null hypothesis in this type of analysis with Friedman is that there are no significant differences among the compared methods. In this case, the null hypothesis can be rejected, since BMSelection and YOLOv8 show apparent differences compared to methods such as RTWAverage or WAFinal.

6 Conclusion

After comparing the results of the YOLOv8 and RT-DETR models in firearm detection in images, it is concluded that YOLOv8 has demonstrated superior performance in several key areas. YOLOv8 not only achieved higher precision (mAP) but also showed better metrics in terms of precision and recall throughout the various training epochs. The model achieved rapid convergence, resulting in a significant reduction in losses (box loss, cls loss, dfl loss) in both the training and validation sets.
On the other hand, RT-DETR, although competitive, did not reach the same levels of efficiency and precision as YOLOv8. RT-DETR's metrics indicated greater variability in precision and recall, as well as slower and less stable convergence compared to YOLOv8.

In summary, YOLOv8 stands out as the most robust and effective option for firearm detection in images within this specific context. Its ability to maintain a high level of accuracy and speed in detection makes it the preferred tool for security applications where both speed and accuracy are crucial.

Experimentation with fusion techniques revealed that the intelligent combination of models can outperform the individual performance of each one in specific scenarios. While direct comparison helps establish a baseline model, fusion strategies—especially the best-result selection—offer a promising path to enhance robustness and accuracy in critical systems such as real-time firearm detection. The choice of the ideal fusion technique will depend on the application context: in a live monitoring system, real-time weighting may be the most effective option; for batch analysis, the classic weighted average may be sufficient.

As firearm detection through deep learning techniques continues to be explored and improved, significant opportunities to advance this area have become evident by integrating new approaches and technologies.

References

Alwakeel, A., Alwakeel, M., Hijji, M., Saleem, T. J., & Zahra, S. R. (2023). Performance evaluation of different decision fusion approaches for image classification. *Applied Sciences, 13*(2), 1059.

DASCI. (n.d.). *Detección de armas*. https://dasci.es/es/transferencia/open-data/deteccion-de-armas/

Hu, Z., Jing, Y., & Wu, G. (2023). Decision-level fusion detection method of visible and infrared images under low light conditions. *EURASIP Journal on Advances in Signal Processing, 2023*.

Müller, D., Soto-Rey, I., & Kramer, F. (2022). *An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks*. *IEEE Access, 10*, 66467. https://doi.org/10.1109/ACCESS.2022.3182399

Othman, N. A., Abdel-Fattah, M. A., & Ali, A. T. (2023). A hybrid deep learning framework with decision-level fusion for breast cancer survival prediction. *Big Data and Cognitive Computing, 7*(1), 50. https://doi.org/10.3390/bdcc7010050

Ouyang, H. (2024). DEYO: DETR with YOLO for end-to-end object detection [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2402.16370

Pereira, D. G., Afonso, A., & Medeiros, F. M. (2015). Overview of Friedman's test and post-hoc analysis. *Communications in Statistics – Simulation and Computation, 44*(10), 2636–2653.

Pereira, L. M., Salazar, A., & Vergara, L. (2024). A comparative study on recent automatic data fusion methods. *Computers, 13*(1), 13. https://doi.org/10.3390/computers13010013

Roboflow Universe. (n.d.). *Pistols dataset*. https://universe.roboflow.com/joseph-nelson/pistols/dataset/1