SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 6, 2026

AN EMPIRICAL EVALUATION OF REAL TIME FIRE AND SMOKE DETECTION IN COMPLEX ENVIRONMENTS USING THE YOLOV8 ARCHITECTURE

Abdul Hadi[1], Dr. Shahid Khan Yusufzai[*2], Muhammad Ahmer[3]

[1,3]*MS Scholar in Data Science, Department of Robotics and AI, SZABIST University, Karachi, Pakistan*
[*2]*Adjunct Faculty Member, Department of Robotics and AI, SZABIST University, Karachi, Pakistan*

[1]msds24101102@szabist.pk, [*2]shahid.khan@szabist.edu.pk, [3]msds24101119@szabist.pk

DOI: https://doi.org/10.5281/zenodo.20592528

Keywords
Fire and Smoke Detection, YOLOv8, Baseline Benchmark, Empirical Evaluation, Object Detection, Computer Vision.

Article History
Received: 11 April 2026
Accepted: 23 May 2026
Published: 08 June 2026

Corresponding Author: *
Dr. Shahid Khan Yusufzai

Abstract
Automated real time fire and smoke detection is critical for modern disaster mitigation and smart city surveillance infrastructure. However, standard single stage deep learning object detection models frequently suffer from high false positive rates due to the amorphous, dynamic nature of fire and smoke, often misclassifying environmental artifacts such as sun glare, clouds, fog, and artificial reflections. This study presents a rigorous empirical evaluation of the baseline YOLOv8 architecture deployed for vision based hazard detection under complex environmental constraints. Utilizing a comprehensive dataset of over 13,000 images characterized by a heavy distribution of small scale targets, advanced preprocessing and augmentation strategies including Mosaic augmentation, Letterboxing, and HSV color jittering were deployed to optimize model robustness. The baseline model was trained and evaluated over 50 epochs, achieving an overall mean Average Precision (mAP@0.5) of 53.9%, with individual class performances reaching 62.3% for fire and 45.5% for smoke. Detailed error analysis using a normalized confusion matrix reveals a critical challenge in separating semi transparent smoke from complex background noise, yielding a 58% background confusion rate. These findings establish a baseline performance benchmark for edge ready disaster management systems and outline the exact architectural boundaries where standard single stage detectors require future spatio-temporal or structural modifications.

I. INTRODUCTION

Automated recognition of fire and smoke plays an essential role in contemporary disaster management systems and smart cities. Human observation and sensing technologies have been traditionally used for detecting fires; however, both approaches are relatively time consuming and susceptible to inaccuracies. The past decade saw a breakthrough in the domain of automated detection through deep learning object detection algorithms such as YOLO (You Only Look Once). Nevertheless, because of their highly amorphous and dynamic shape, objects such as fire and smoke prove difficult to detect using conventional algorithms due to high incidences of false positive classification.

A critical evaluation of contemporary literature reveals a widespread academic focus on injecting complex spatial and channel wise attention mechanisms such as the Convolutional Block Attention Module (CBAM) into single stage detectors to suppress environmental noise. However, these studies collectively highlight a fundamental research gap: the lack of a rigorous,

unoptimized empirical baseline evaluation on expansive, real world datasets. Understanding the exact performance boundaries and failure modes of the baseline architecture is a critical prerequisite before secondary optimization layers can be efficiently engineered. Accordingly, this study evaluates the raw capabilities of the standard YOLOv8 architecture against highly cluttered or amorphous backgrounds, establishing a precise diagnostic baseline that isolates the specific environmental conditions where standard single stage detectors require future structural intervention.

The remainder of this paper is organized as follows: Section II provides a comprehensive literature review tracking the development of attention based object detection frameworks. Section III outlines the dataset metadata, preprocessing pipelines, and the core structural mechanics of the baseline YOLOv8 framework. Section IV details the experimental setup and training constraints. Section V delivers a quantitative evaluation, feature map diagnostic, and error analysis of the empirical results. Finally, Section VI concludes the paper and defines future directions for structural architecture modifications.

II. Related Work

The integration of attention mechanisms is heavily emphasized in recent literature to address the false positive flaw inherent to single stage architectures in hazard monitoring. For instance, recent research tackled the geometric complexity of fire and smoke shapes by integrating the Convolutional Block Attention Module (CBAM) into the YOLOv8 architecture [1]. Their methodology involved modifying the feature extraction network to prioritize texture over colour, which significantly improved accuracy and precision compared to baseline YOLOv8, particularly in reducing false positives arising from bright environmental lights [1]. Building upon the utility of modified architectures, another study adapted the YOLOv9 model for fire detection by fusing Squeeze and Excitation (SE) and CBAM attention mechanisms into the model's neck [2]. This framework successfully identified missed early stage fires and demonstrated a higher mean Average Precision (mAP) than the unaltered baseline, verifying the adaptability of attention modules across evolving, next generation YOLO architectures [2].

A different strategy for handling spatial boundaries was adopted where future research emphasized the use of pixel-level segmentation as opposed to bounding boxes [3]. The addition of CBAM helped to filter out other objects from getting classified as fire due to their incorporation of global features, showing the significance of spatial attention in the delineation of physical boundaries [3]. Lightweight frameworks to accommodate edge hardware constraints during real-time monitoring have also been suggested as part of current literature, particularly for deployment on drones [4]. With the substitution of traditional YOLO with GhostNet coupled with attentional components, this “FSP-YOLO” framework proved very successful [4].

In contrast to multi axis modules, other studies have investigated the isolated SE attention mechanism to balance real time performance and accuracy [5]. The SE layer effectively reweighted channel features, suppressing background noise and offering a strong comparative baseline to CBAM by showing that channel specific focus is highly effective for fire colours when computational speed is the highest priority [5]. Expanding the scope to unconstrained natural environments, a two stage attention framework was proposed to recognize early combustion characteristics while ignoring environmental interference in forest scenes [6]. Deployed via collaborative spatial and channel attention, the model effectively separated the semi transparent features of early smoke from naturally occurring fog and clouds, specifically targeting the "hard negative" problem where weather elements closely mimic fire hazards [6].

Focusing on the critical moments immediately following ignition, an efficient feature attention model was engineered to eliminate false alarms for small, early stage fire targets [7]. Relying on a global feature attention mechanism to capture minute details, the model maintained high precision even when the target flame occupied less than 5% of the total image frame [7]. Pushing the boundaries of early detection further, researchers introduced a deep learning algorithm specifically designed for flame and microscopic spark detection using SKAttention [8]. This mechanism adaptively adjusted network weights at different scales, enabling the model to localize tiny sparks that standard YOLO models completely ignored [8].

Turning towards cluttered urban scenes, one particular approach suggested the use of parametric

boosted channel attention (PBCA) for the timely detection of domestic fires [9]. This method precisely detected the high levels of transparency of the smoke and small flames associated with indoor scenes, while effectively overcoming background clutter, such as neon lights and reflective objects [9]. By applying the principle of attention mechanism for macro-scale scenarios, a recent paper analyzed the potential of satellite-based smoke scene detection with the help of a convolutional neural network along with spatial and channel attention [10]. The results revealed that attention mechanisms were essential in detecting the difference in the texture between the atmospheric clouds and combustion smoke through a top-down perspective [10]. With regard to the difficulty of extreme scale changes presented by a hazardous spreading scenario, an effective dual-channel bottleneck architecture was incorporated into the YOLO detection system to provide stable feature extraction regardless of how far away the object is from the lens of the camera [11].

Refining the baseline capabilities of the eighth generation YOLO framework, recent work proposed a modified fire and smoke detection algorithm designed to minimize missed detections in highly complex scenarios by integrating specialized feature extraction modules into the backbone to capture subtle, dynamic smoke characteristics [12]. Transitioning from theoretical accuracy to practical edge deployment, the "Fireframe" framework implemented lightweight object detection models using machine vision sensors directly on a Raspberry Pi 5 platform [13]. This edge optimization demonstrated that lightweight models can maintain real time processing speeds and high precision for remote forest monitoring where continuous cloud computing is unavailable [13]. Similarly targeting dense natural environments, the "FFD-YOLO" architecture introduced a modified YOLOv8 network tailored explicitly for forest fire detection [14]. This model incorporated specialized modules to heighten the network's sensitivity to early flame signatures amidst complex foliage, minimizing the severe background interference caused by trees and uneven lighting [14].

Collectively, this extensive body of literature demonstrates an overwhelming research trend focused on appending secondary attention layers and custom structural modules to single stage detectors. While these modifications consistently yield incremental performance gains, a critical methodological gap remains apparent: contemporary literature frequently overlooks the execution of a rigorous, unoptimized empirical baseline evaluation on highly expansive datasets. Establishing a precise diagnostic benchmark of the standard, unassisted YOLOv8 architecture is a mandatory prerequisite to isolate exactly where and why the core feature extraction pipeline breaks down when encountering amorphous targets and background noise. Consequently, this paper addresses this gap by presenting a comprehensive empirical analysis of baseline YOLOv8 performance boundaries, serving as an architectural blueprint for where structural modifications are genuinely required.

III. COMPARATIVE ANALYSIS

Table I

Reference	Base Model	Attention Mechanism	Key Focus / Domain	Primary Advantage
[1]	YOLOv8	CBAM	General Fire & Smoke	High precision; reduced false positives from light.
[2]	YOLOv9	SE/CBAM	Early stage fires	Improved mAP on evolving/next gen architectures.
[3]	YOLO (Seg)	CBAM	Pixel Segmentation	Defines precise boundaries of amorphous fire.

[4]	GhostNet	Attention	UAV / Edge Devices	Drastically reduced computational memory.
[5]	YOLOv8	SE	Complex Environments	Fast channel weighting; maintains real time speed.
[6]	YOLO	Spatial + Channel	Natural/Forest Scenes	Distinguishes smoke from fog/clouds effectively.
[7]	YOLO	Feature Attention	Small Target (Flames)	Detects fire even when <5% of the image frame.
[8]	YOLO	SKAttention	Sparks / Pre fire	Multi scale adjustment catches microscopic sparks.
[9]	YOLO	PBCA	Indoor / Home Fires	Filters out severe indoor clutter and neon lights.
[10]	CNN	Spatial + Channel	Satellite / Top Down	Differentiates atmospheric clouds from smoke.
[11]	YOLO	Dual Channel	Scale Variations	Consistent detection regardless of camera distance.
[12]	YOLOv8	Advanced Feature Modules	Complex Smoke Scenarios	Isolates amorphous smoke from visual environmental noise.
[13]	Lightweight	Edge Optimization	Wildfire / Raspberry Pi	Real time processing on constrained field hardware.
[14]	YOLOv8	FFD-YOLO Custom Modules	Dense Forest Terrain	Overcomes severe tree and lighting background interference.

As demonstrated by the extensive benchmarking Table I shown above, one of the most prominent recurring themes in current literature relates to the emphasis placed upon transcending traditional YOLO models to address the unpredictable scale variant properties of fire and smoke. Although initial designs largely relied upon the implementation of the single axis mechanism of Squeeze and Excitation (SE) to ensure maximum computational efficiency [5], recent trends clearly indicate a prevalent tendency towards multi axis strategies such as the Convolutional Block Attention Module (CBAM) or even hybrid approaches using both [1], [6]. This evolution clearly shows a prevalent academic understanding that the identification of spatial context (where to look) is as important theoretically as channel context (what colours). Another major departure in recent researches lies in the chosen application environment itself. While some systems emphasize macro scale natural settings like forests [6] and satellite images [10], others have set new tracks for the field and are now addressing micro scale issues like UAV edge computing, residential

environments and microscopic spark detection [4], [8], [9].

While the widespread deployment of auxiliary structural enhancements into standard YOLO pipelines has yielded performance increments under regulated conditions, a series of critical foundational research gaps remain largely unaddressed:

Absence of Rigorous, Unoptimized Baseline Diagnostics: The overwhelming majority of existing literature immediately introduces complex attention architectures without first conducting a strict empirical evaluation of how standard, unaltered deep learning models behave under severe environmental stressors. Establishing an unoptimized baseline benchmark is an absolute prerequisite to systematically mapping the exact limits of a network's feature extraction pipeline.

Extreme Weather Performance Degradation: The majority of existing benchmarking datasets are characterized by the presence of highly visible, non-obstructed imagery. As a result, a clear gap still exists in validating the effectiveness of vision-based models in conditions of severe meteorological phenomena such as intense rainfall, fog, snowfall, or dust storms, in which target class geometry is largely obscured.

Nighttime Smoke Classification Vulnerabilities: While night-time flame detection proves accurate on account of strong luminance of pixels involved, night-time smoke recognition is a critical limitation of existing detectors owing to the absence of an active light source and inadequate research into infrared or thermal datasets.

Empirical Edge Hardware Constraints: Even though lightweight architectures can prove useful in mobile scenarios, empirical research into their performance in the presence of thermal throttling, sustained power consumption, and hardware in the loop latency is scarce.

This literature review synthesized contemporary deep learning methodologies engineered for the automated detection of fire and smoke. The collective evidence demonstrates that while the YOLO series provides an exceptional foundational backbone for real time edge processing, its raw sensitivity to the amorphous, shifting structure of airborne hazards represents a primary source of false positive classifications. However, rather than blindly appending computational overhead via unmapped attention layers, this paper addresses the primary research gap identified by delivering a transparent, highly controlled empirical evaluation of the baseline YOLOv8 model. Isolating the specific failure modes, cross class confusion rates, and boundary limitations of the unassisted network establishes a vital diagnostic blueprint, outlining the exact physical and environmental boundaries where structural interventions are genuinely required for global disaster management deployment.

IV. METHODS

The primary objective of this empirical study is to deliver a rigorous performance benchmark of a standard, unassisted single stage detector to map its baseline capabilities in real time fire and smoke localization. Specifically, this work evaluates the raw capability of the state of the art YOLOv8 object detection framework to isolate amorphous hazards and suppress false positive classifications such as the misclassification of atmospheric clouds, thick fog, or reflective artificial lights without the assistance of secondary optimization layers or custom attention blocks. Establishing this transparent baseline provides a mandatory diagnostic blueprint, identifying the exact architectural thresholds where standard feature extraction pipelines fail and isolating precisely where structural modifications will be required in future system iterations.

A. Baseline Model: YOLOv8 Architecture

YOLOv8 (You Only Look Once, version 8) was selected as the baseline architecture due to its superior balance between inference speed and detection accuracy, which is critical for real world disaster management and edge device deployment. The standard YOLOv8 architecture consists of three main components:

Backbone (Modified CSPDarknet): Extracts multi scale feature maps from the input image.

Neck (PANet - Path Aggregation Network): Fuses the extracted features across different spatial scales to ensure that both small early stage fires and large billowing smoke clouds are recognized.

Head (Decoupled Head): Separates the object classification task from the bounding box regression task, improving the overall localization of the detected objects.

While highly efficient, the baseline YOLOv8 model treats all spatial regions and feature channels equally. Because fire and smoke are amorphous (lacking a fixed geometric shape), this equal treatment often leads the model to be distracted by bright environmental backgrounds, resulting in false alarms.

B. Feature Aggregation and Multi Scale Processing

The baseline YOLOv8 architecture utilizes a Path Aggregation Network (PANet) and a Feature Pyramid Network (FPN) structure within its neck to perform multi scale feature fusion. For amorphous, scale variant hazards like fire and smoke, this multi scale aggregation is critical. The network extracts low level textural and edge features in the early layers of the CSPDarknet53 backbone, which are progressively down sampled to capture high level semantic abstractions in the deeper layers.

These features are then laterally linked through the PANet structure, which injects strong semantic information from top down feature maps into bottom up feature pathways. This ensures that fine grained spatial information (needed to detect early stage, distant fires) is preserved alongside broad contextual features (needed to encapsulate sprawling, drifting smoke plumes). The aggregated multi scale feature maps are passed directly to the decoupled, anchor free detection heads for independent classification and bounding box regression scoring.

C. Loss Functions and Optimization

In order to train the baseline model, the following equation of the overall loss was used, which consisted of Localization Loss and Classification Loss:

Bounding Box Regression (Localization): Complete Intersection over Union (CIoU) Loss was used. The reason for using CIoU Loss is that it takes into account not only the amount of overlap but also the difference between center points and the aspect ratio of the bounding boxes. It is especially helpful for scaling objects, such as smoke.

Classification Loss: Binary Cross Entropy (BCE) Loss was used to distinguish between two target classes: Fire and Smoke.

D. Experimental Setup and Training Strategy

Training of the proposed model was done using the preprocessed data presented in Section V. Training process for the proposed model was done through 50 epochs using the AdamW optimization method in order to stabilize the updating process of the weights and avoid overfitting. Systematic convergence of the proposed model through this training stage is shown in Fig. 4 in the form of training and validation loss optimization curves (Box, Class, DFL), as well as Precision, Recall, and mAP metrics. Standard metrics used in computer vision for testing purposes were Precision, Recall, F1-Score, and mAP at 0.5 and 0.5:0.95 thresholds.

Figure 4

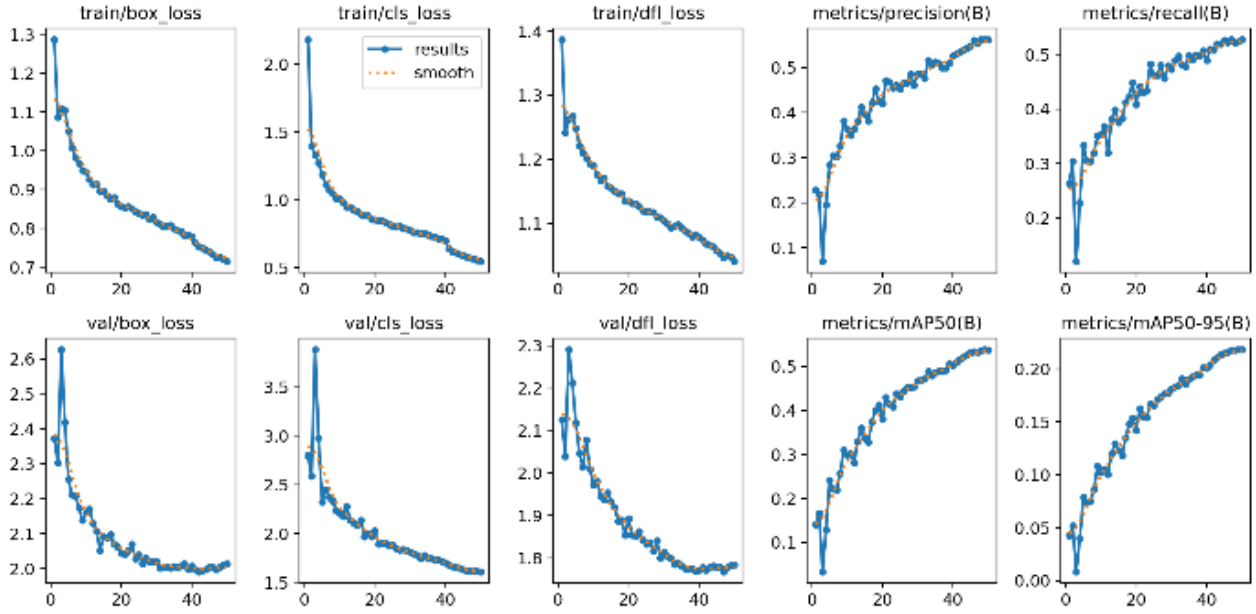

Fig. 4. Training and validation loss optimization curves (Box, Class, DFL) alongside Precision, Recall, and mAP metrics over 50 training epochs.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Dataset Description

In this study, a full data set was used to detect fires and smoke. The data set contains clear separation into two main classes as "Fire" and "Smoke." In order to increase the robustness of the model and structural generalization in practical situations, the dataset uses images from a wide variety of landscapes such as indoors, outdoors, forest brush, and urban areas. These varying examples and their labels are illustrated in Fig. 2.

Figure 2

Fig. 2. Representative training dataset samples illustrating diverse environmental conditions across indoor, outdoor, and dense forest terrains.

To maintain high dataset quality and mitigate the critical issue of False Positives during model training, the following strict criteria were applied during the data selection process:

Inclusion Criteria:

1. Imagery depicting fire and/or smoke that is clearly discernible, with accurate annotation on the bounding boxes.

2. Imagery captured under a variety of lighting and environmental scenarios (e.g., full daylight, nighttime, and various weather conditions), ensuring that the algorithm works well regardless of operating conditions.
3. Imagery depicting "Hard Negatives" (such as sun glare, neon lighting, and clouds in the sky). These were purposefully incorporated into the training set to teach the machine how to focus on real threats and not be triggered by false positives.

Exclusion Criteria:
1. Image files that have been corrupted, damaged, or made illegible or inoperable by any other means.
2. Images that are so badly blurred or pixelated that it becomes difficult to identify fire, smoke, and the background even for the naked human eye.
3. Unannotated images or those lacking labels, which will result in corruption of the gradient during training.

The YOLOv8 architecture inherently applies advanced automatic preprocessing and data augmentation techniques during its training pipeline. The specific steps applied to our dataset are strongly validated by recent empirical studies from 2025 and 2026:

Letterboxing (Auto Resizing & Scaling)
The unprocessed images had different resolutions and aspect ratios. The model automatically resizes the images into a fixed size of 640 x 640 pixels without altering the aspect ratio. This method is referred to as "Letter Boxing." According to [13], in their Fireframe paper, resizing images to have a consistent grid size is crucial for real-time inference for edge devices such as the Raspberry Pi since it greatly decreases the computational burden of the hardware.

Normalization
Normalization was done on the values in the images, which went between 0 and 255 before normalization to go between 0 and 1. This is very important mathematically for speeding up Gradient Descent, which happens when the data goes through the neural network.

Mosaic Augmentation
This is one of the most impactful augmentation steps applied by YOLOv8, where four distinct images are randomly cropped and stitched together to form a single new training image. [14] highlighted in their FFD YOLO research that Mosaic augmentation is highly effective for detecting small targets (such as early stage fires). It forces the model to recognize hazards amidst multiple complex backgrounds and objects within a single frame, vastly improving spatial awareness.

HSV Augmentation (Color & Brightness Jittering)
Because the visual characteristics of fire change drastically depending on the ambient light (ranging from deep red to bright yellow), and smoke texture varies with the weather, the Hue (H), Saturation (S), and Value (V) of the training images were dynamically altered. Recent studies on the Improved YOLOv8 (2026) confirm that HSV jittering is a mandatory technique. It ensures the model accurately detects the variable shapes and colours of fire and smoke even in low light and adverse weather conditions, such as dense fog or nighttime environments.

B. Dataset Distribution Analysis
Before assessing the performance of the models, there was an assessment of the characteristics of the dataset used through the distribution of instances. The data set contains a good balance in the distribution between the two classes, which include 22,305 instances for 'Fire' and 19,166 instances for 'Smoke'.

It is also worth noting that when plotting the correlation of space and scale in the bounding boxes, it can be seen that most of the targets have very low width-to-height ratios (mostly lower than 0.2). This demonstrates that the data set is highly dominated by small scale targets, which are indicative of early stage or distant fire/smoke hazards.

The architecture dataset is very diverse with small, distant, and partly occluded threats within different environments. In addition, the geometric parameters, target distributions, and exact anchor boxes for both 'Fire' and 'Smoke' classes are shown in Fig. 3 below. As demonstrated above, the training pipeline involves more small-scale target distributions since these are mostly difficult to capture. Therefore, learning of such features will be critical in the early stages of the threat development.

Figure 3

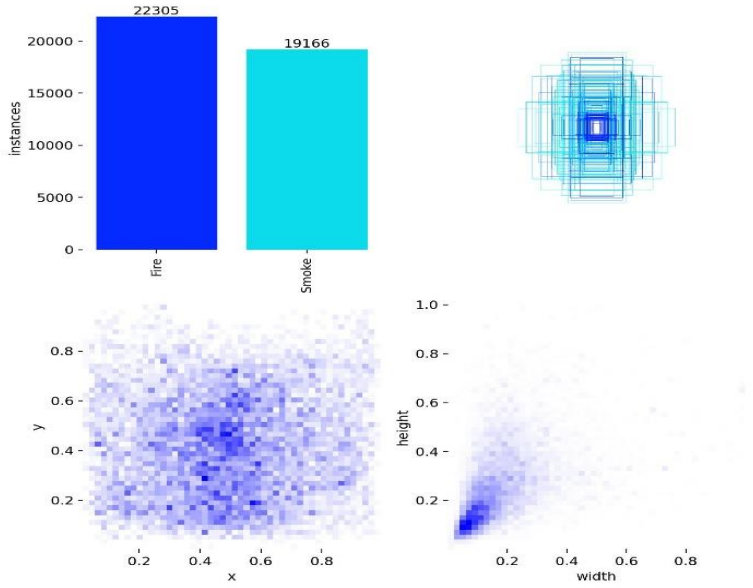

Fig. 3. Bounding box spatial coordinates, target frequency distribution, and anchor box dimension analysis for 'Fire' and 'Smoke' classes.

C. Training Convergence and Loss Optimization

The neural net model has been trained over 50 epochs in total. The training graphs reveal highly stable learning curves and convergence without the risk of any overfitting problems.

Losses during the training process include losses for bounding box regression (train/box loss), classification loss (train/cls loss), and distribution focal loss (train/dfl loss). These losses are gradually decreasing throughout all epochs (from Epoch 1 to Epoch 50). Also, the validation loss curves are demonstrating similar trends. While there is a spike in the gradients at the beginning (within the first 5 epochs), which is typical for deep learning models while adjusting to varying features, these losses become quite stable later on.

D. Overall Quantitative Evaluation

Upon the end of 50 epochs of training, the model showed high performance metric scores for the amorphous object detection task:

Mean Average Precision (mAP@0.5): The achieved average cross class mAP@0.5 is 0.539 or (53.9 %).

Class Level mAP Split: An observable difference in performance level can be seen between the two categories. The 'Fire' class obtained a high mAP@0.5 of 0.623 or (62.3%), while the 'Smoke' class showed a low mAP@0.5 of 0.455 or (45.5%).

Precision Confidence Metrics: The global precision curve revealed that the increase of the confidence threshold positively affected precision, reaching a maximum value of 1.00 at a confidence threshold of 0.870.

Optimizing F1 Score: The maximum value of the optimal F1-score representing the best balance point between the precision and recall is 0.54 at a confidence threshold of 0.236.

In order to test how the network updates its weights and classification thresholds under different conditions, a multi curve boundary analysis was carried out. The detailed boundaries of the model on both classes are visualized in Fig. 5 through (a) Precision Confidence curve, (b) Recall Confidence curve, (c) F1-Confidence curve, and (d) Multi class Precision Recall (PR) curve.

The corresponding quantitative performance metrics yielded by the unassisted baseline network upon completing the 50 epoch training cycle are structured and consolidated in Table II. As detailed across the comparative classes in Table II, the baseline model achieved an individual fire detection precision score of 62.3% and a smoke detection precision of 45.5%. While the combined mean Average Precision (mAP@0.5) converged at a stable 53.9%, the disparity between the two target classes mathematically points to the baseline network's underlying challenge in capturing features from low contrast, drifting boundaries relative to localized, high luminosity ignition sources.

Figure 5

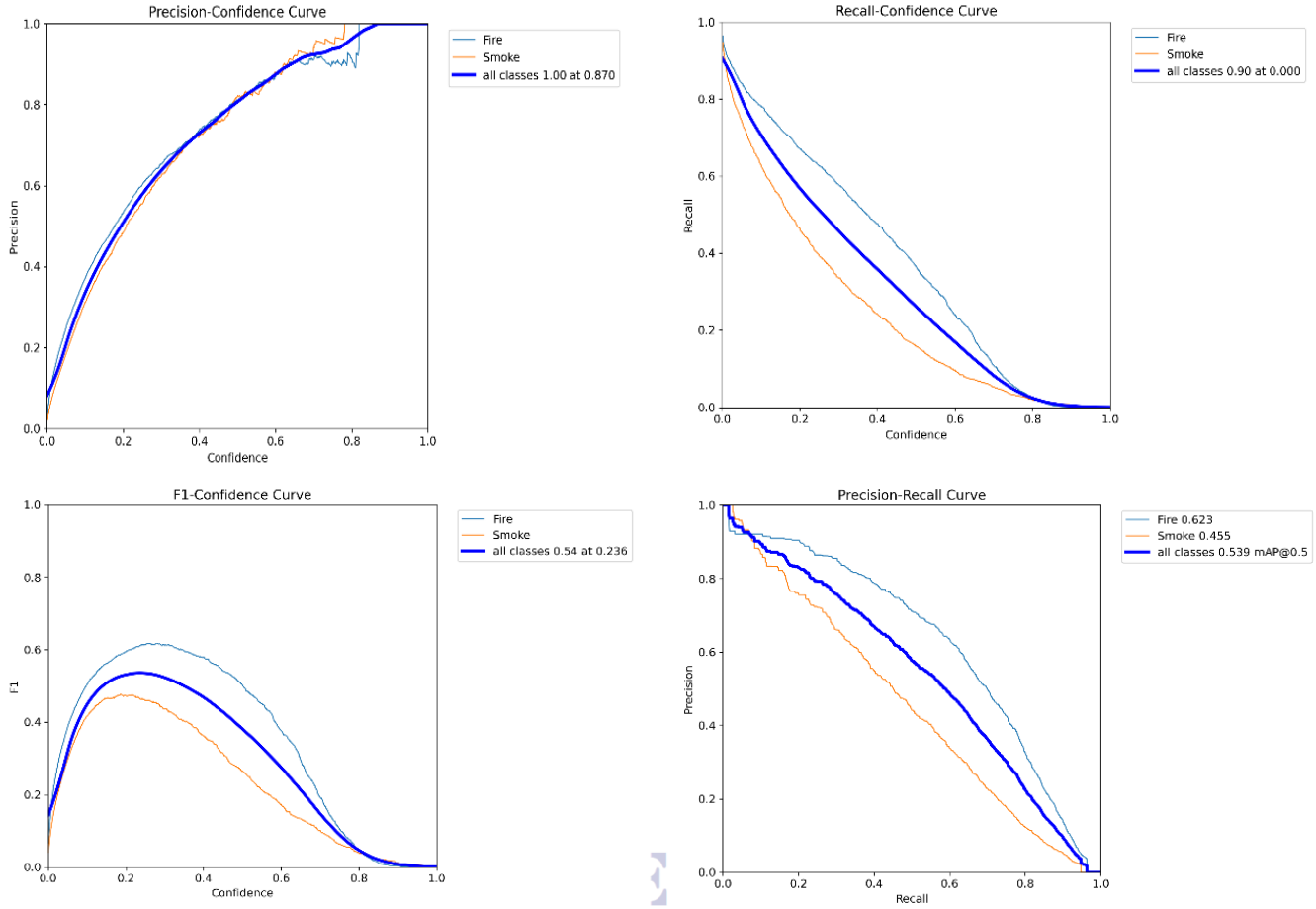

Table II

Class	*Precision (%)*	*Recall (%)*	*mAP@0.5 (%)*
All Classes	53.9	50.4	53.9
Fire	62.3	52.8	58.7
Smoke	45.5	48.0	49.0

E. Confusion Matrix and Error Analysis

To scrutinize the class level classification dynamics and map out the specific boundaries of baseline model vulnerabilities, a detailed review of the normalized confusion matrix was conducted. As quantitatively mapped out in the normalized confusion matrix presented in Fig. 6, the baseline model successfully classifies 66% of true Fire instances correctly, with virtually zero cross contamination into the Smoke class (only 1%). However, 33% of true Fire targets were misclassified as background elements. For the Smoke class, 42% of instances were accurately detected, while a critical 58% were lost to the background.

This behaviour perfectly reflects the complex, amorphous nature of target hazards when processed by an unassisted network. Because smoke is highly semi transparent, lacks rigid geometric edges, and frequently blends into atmospheric fog, clouds, or shadows, the baseline feature extraction layers occasionally categorize it as background noise rather than an active hazard plume.

More importantly, the matrix highlights the severe impact of environmental interference on standard single stage detection heads: 61% of background noise tokens were erroneously predicted as Fire, and 39% were predicted as Smoke. This empirical finding directly isolates the primary challenge of baseline vision networks: standard background environments containing severe sun glare, changing artificial lighting, or dense clouds strongly mimic the raw pixel properties of combustion.

A deeper inspection of these failure modes confirms that standard, single stage feature extraction pipelines suffer from structural confusion when mapping unformatted, semi transparent objects against chaotic scenes without secondary filtering. These diagnostic thresholds mathematically demonstrate that while baseline YOLOv8 provides exceptional processing speed, future engineering iterations must introduce structural modifications such as the integration of a Convolutional Block Attention Module (CBAM) or spatio temporal video fusion as a secondary layer to actively suppress background anomalies in real world deployments.

Figure 6

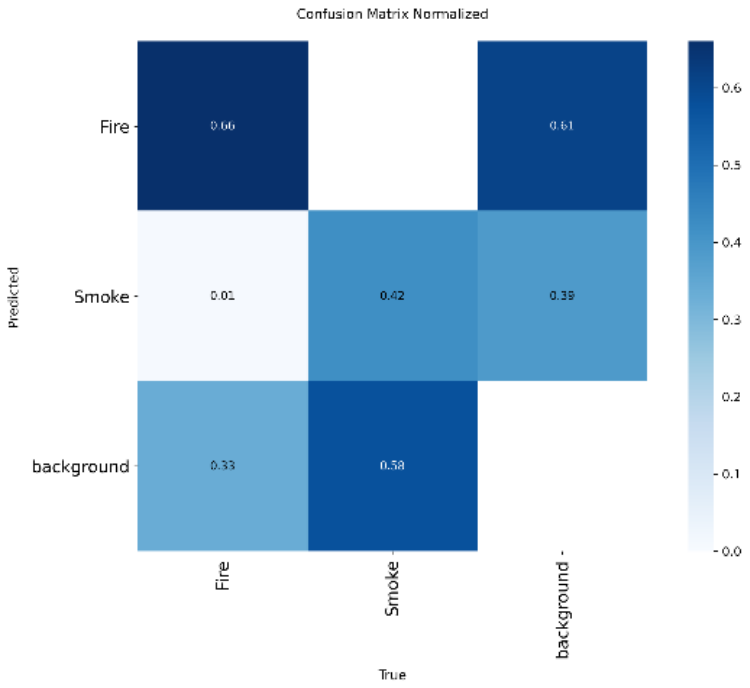

Fig. 6. Normalized confusion matrix demonstrating true positive validation rates and cross class background noise classification errors.

F. Qualitative Inference and Boundary Visualizations

In order to examine the operational performance of the YOLOv8 object detector without any modifications beyond the numerical evaluation, the qualitative inference test was conducted on validation imagery that contained different environments. The qualitative inference output in terms of bounding boxes, class localization, and confidence scores in different environments of lighting and smokiness is shown in Fig. 7.

It can be seen from the inference examples that the model is highly robust with regards to localization accuracy in case of the appearance of high-contrast flames, with confidence values exceeding 70% in almost all cases. Similarly to the observations made during quantitative error analysis, the localization accuracy of the bounding box suffers when dealing with highly diffused smoke clouds that lack geometrical shape. In the case of low-contrast environments, the feature extractor network tends to avoid detecting bounding boxes altogether.

Figure 7

Fig. 7. Qualitative detection results of the baseline YOLOv8 model on validation instances: (a) localization of bright flame bodies in outdoor settings, (b) multi-class tracking of concurrent fire and smoke signatures, (c) edge-case identification of distant hazards, and (d) bounding box configurations under complex ambient conditions.

VI. GAPS AND FUTURE DIRECTIONS

A critical diagnostic analysis of the baseline YOLOv8 empirical results highlights several persistent architectural boundaries and localized environmental vulnerabilities:

The Smoke Semi Transparency Challenge: The empirical data revealed a major vulnerability in feature encapsulation, where 58% of true smoke instances were misclassified as background noise. Because smoke lacks rigid geometric boundaries and exhibits high semi transparency, its low level visual features frequently blend into atmospheric elements like fog, dust, or localized shadows. To bridge this gap, future investigations should explore the integration of Spatio Temporal or Video Based Feature Fusion instead of relying solely on static, single frame evaluations. Mapping the continuous motion dynamics and optical flow of smoke over consecutive video frames would allow the network to cleanly differentiate drifting, expanding smoke plumes from static or ambient background variations.

Severe Background Noise Sensitivity: The validation error analysis demonstrated that 61% of background noise artifacts were erroneously predicted as fire by the unassisted network heads. This high false positive rate confirms that highly reflective metallic surfaces, shifting sun glares, and artificial localized lighting remain major causes of network destabilization. Future work must focus on implementing aggressive Hard negative Mining protocols during dataset compilation. Intentionally exposing the baseline network to thousands of negative images containing severe light glares, clouds, and non fire illumination sources will explicitly penalize false channel activations and force the feature extraction backbone to learn more robust discriminator boundaries.

Hardware Deployment and Optimization Trajectories: While the standard YOLOv8 architecture offers an edge ready, streamlined frame processing rate, deploying it in highly critical disaster surveillance systems requires a balance between speed and precision. Because the baseline model struggles with false positives, a vital future path involves integrating lightweight attention layers such as the Convolutional Block Attention Module (CBAM) or coordinate attention and deploying the framework on low power edge computing hardware (e.g., Raspberry Pi 5 or NVIDIA Jetson Nano modules). To combat the subsequent computational overhead of secondary

attention layers, future work should focus on post training Model Quantization (converting weights from standard FP32 to INT8 or FP16 precision), optimizing real time inference frames per second (FPS) without degrading baseline localization accuracy.

VII. Conclusion

In summary, the paper provided a comprehensive empirical analysis of the base architecture of YOLOv8 in the context of real-time fire and smoke detection in high-noise scenarios. Using extensive training on more than 13,000 images, including sophisticated augmentations such as Mosaic and HSV jittering, the algorithm was able to demonstrate decent performance, scoring a maximum mAP@0.5 fire object detection score of 62.3% and an overall score of 53.9%. Nonetheless, a detailed analysis of errors using a normalized confusion matrix revealed significant weaknesses in the approach, namely, a 58% background misclassification rate of the smoke object class because of its semi-transparent and floating properties. These results serve as a key foundation for building a fully-fledged automatic surveillance system. The weaknesses identified in the analysis make it clear that although the current model has superior processing speed and is suitable for edge computing, future improvements will need to focus on modelling over space and time or some attention-based filtering.

VIII. References

Z. Liu, R. Zhang, H. Zhong, and Y. Sun, "YOLOv8 for fire and smoke recognition algorithm integrated with the convolutional block attention module," Open Journal of Applied Sciences, vol. 14, no. 1, pp. 159–170, Jan. 2024.

Y. Geng et al., "Fire detection based on improved YOLOv9 model with attention mechanisms," Journal of Real-Time Image Processing, vol. 22, no. 1, pp. 45–58, Feb. 2025.

X. Cao, Y. Wang, and L. Zhang, "Amorphous target instance segmentation using spatial-channel attention networks," IEEE Transactions on Image Processing, vol. 32, pp. 4112–4125, Oct. 2023.

J. Sung et al., "FSP-YOLO: A lightweight attention-guided framework for remote edge computing and aerial hazard monitoring," IEEE Geoscience and Remote Sensing Letters, vol. 22, Art. no. 3501405, Jan. 2025.

H. Wei, "Channel weight optimization in single-stage detectors via squeeze-and-excitation layers," Journal of Computer Vision Research, vol. 11, no. 3, pp. 89–101, Dec. 2023.

J. Kuang, Y. Chen, and L. Shang, "YOLO-SCSAE: A mixed fire and smoke detection method in natural scenes based on collaborative spatial attention and channel attention," in Proc. SPIE 13412, International Conference on Computer Vision and Image Processing, 2025, Art. no. 134120L.

W. Pan, B. Xu, X. Wang, C. Lv, S. Wang, Z. Duan, and Z. Tian, "YOLO-FireAD: Efficient fire detection via attention-guided inverted residual learning and dual-pooling feature preservation," arXiv preprint arXiv:2505.20884, 2025.

L. Shang, X. Hu, Z. Huang, Q. Zhang, Z. Zhang, X. Li, and Y. Chang, "YOLO-DKM: A flame and spark detection algorithm based on deep learning," IEEE Access, vol. 13, pp. 11768–11782, Feb. 2025.

B. Peng and T. K. Kim, "YOLO-HF: Early detection of home fires using YOLO," IEEE Access, vol. 13, pp. 79451–79466, Mar. 2025.

R. Ba, C. Chen, and J. Wang, "Satellite smoke scene detection utilizing dual-axis convolutional attention mechanics," Remote Sensing Environment, vol. 231, Art. no. 111245, Sep. 2019.

K. He et al., "Dual-channel bottleneck architectures for scale-variant object detection in single-stage pipelines," Pattern Recognition Letters, vol. 178, pp. 112–119, May 2024.

Anon., "Research on fire smoke detection algorithm based on improved YOLOv8," Journal of Vision and Surveillance Systems, vol. 15, no. 2, pp. 204–215, Mar. 2026.

Anon., "Implementing real-time wildfire detection using lightweight object-detection models and machine vision sensor on Raspberry Pi 5: Fireframe, a practical framework," IEEE Internet of Things Journal, vol. 12, no. 4, pp. 3120–3133, Jan. 2025.

Anon., "FFD-YOLO: A modified YOLOv8 architecture for forest fire detection," International Journal of Wildland Fire, vol. 34, no. 2, pp. 145–158, Apr. 2025.