Libyan Open University Journal of Applied Sciences (LOUJAS)
Volume 2 - Issue 2 - 2026 - Pages 220-230

Early Wildfire Detection Using YOLOv8 Nano Lightweight Classification

Fadwa Fawzi Al Tawati Al Manfi*
Department of Computer science , Libyan Academy for Graduate Studies, Derna Branch, Libya
***Email (for reference researcher):** fadwa@lad.edu.ly

الكشف المبكر عن حرائق الغابات بالاعتماد على نموذج YOLO v8

فدوى فوزي التواتي المنفي
قسم الحاسوب، الاكاديمية الليبية للدراسات العليا فرع درنة ،ليبيا

Received: 15-07-2026; Accepted: 01-09-2026; Published: 05-09-2026

Abstract:
The development of rapid, portable early detection technologies is necessary due to the increasing frequency and intensity of wildfires. In order to detect wildfires, this paper suggests a lightweight binary classification framework based on YOLOv8 microcls for edge devices with limited resources. The model achieves 96.02% overall accuracy on a test set of 402 images (156 fire, 246 no fire), with per-class F1 scores of 94.84% (fire) and 96.76% (no fire), after training on the Wildfire Dataset (2n version) for 30 epochs at 320x320 resolution. Three primary failure modes are identified by a thorough error analysis of the sixteen incorrectly classified samples: misreading of distant little fires, smoke-occluded flames (44.4% of false negatives), and sunset/sunrise chromatic confusion (42.9% of false positives). An analysis of public data for YOLOv5n, EfficientNet B0, ResNet 50, and MobileNetV2 assessed on data for ResNet 50, MobileNetV2, EfficientNet B0, and YOLOv5n, each assessed on its own original dataset, positions YOLOv8n cls favorably on the accuracy-efficiency tradeoff. With only 3.9 million parameters and inference speeds exceeding 100 FPS on GPU, the proposed system is architecturally suitable for real-time deployment on Raspberry Pi and IP camera networks, pending direct edge hardware validation. Targeted mitigation strategies, including dynamic thresholding, multi-frame temporal consistency, and seasonal domain adaptation, are proposed to reduce the false negative rate below 3% and the false positive rate below 2%.

Keywords: Wildfire detection, YOLOv8, lightweight CNN, edge computing, binary classification, error analysis, domain adaptation, real-time inference.

الملخص:
نظرًا للتكرار والشدة المتزايدَين لحرائق الغابات، أصبح تطوير تقنيات كشف مبكر سريعة ومحمولة أمرًا ضروريًا. تقترح هذه الورقة، من أجل الكشف عن حرائق الغابات، إطار عمل خفيف الوزن للتصنيف الثنائي قائم على YOLOv8 microcls مخصص للأجهزة الطرفية ذات الموارد المحدودة. يحقق النموذج دقة إجمالية بلغت 96.02% على مجموعة اختبار مكوّنة من 402 صورة (156 صورة حريق و246 صورة بلا حريق)، مع درجات F1 لكل فئة بلغت 94.84% (للحريق) و96.76% (لعدم الحريق)، وذلك بعد التدريب على مجموعة بيانات الإصدار 2 لمدة 30 حقبة تدريبية بدقة 320×320 بكسل.
حدد تحليل دقيق للأخطاء، أُجري على العينات الست عشرة المصنَّفة تصنيفًا خاطئًا، ثلاثة أنماط رئيسية للفشل، وهي: سوء قراءة الحرائق الصغيرة البعيدة، والألسنة النارية المحجوبة بالدخان (44.4% من الحالات السلبية الخاطئة)، والالتباس اللوني الناتج عن غروب/شروق الشمس (42.9% من الحالات الإيجابية الخاطئة).
كما أن تحليل بيانات عامة لكل من YOLOv5n و EfficientNet B0و ResNet 50و MobileNetV2 — حيث قُيِّم كل نموذج على مجموعة بياناته الأصلية الخاصة به — يضع نموذج YOLOv8n cls في موقع مُفضَّل من حيث التوازن بين الدقة والكفاءة. ونظرًا لامتلاكه 3.9 مليون معلمة فقط، وسرعات استدلال تتجاوز 100 إطار في الثانية على وحدة معالجة الرسوميات(GPU) ، فإن النظام المقترح مناسب من الناحية المعمارية للنشر في الزمن الحقيقي على أجهزة Raspberry Piوشبكات كاميراتIP ، وذلك ريثما يتم التحقق المباشر من أدائه على العتاد الطرفي.

وأخيرًا، تُقترح استراتيجيات تخفيف مستهدفة، من بينها العتبة الديناميكية ، والاتساق الزمني متعدد الإطارات ، والتكيّف المجالي الموسمي ، من أجل خفض معدل الحالات السلبية الخاطئة إلى أقل من 3%، ومعدل الحالات الإيجابية الخاطئة إلى أقل من 2%.

الكلمات المفتاحية: كشف حرائق الغابات، ، الشبكات العصبية التلافيفية خفيفة الوزن ، الحوسبة الطرفية ، التصنيف الثنائي، تحليل الأخطاء، التكيّف المجالي، الاستدلال في الزمن الحقيقي

I. Introduction

Because of climate change, extended droughts, and human activity, wildfires are now among the most destructive natural disasters of the twenty-first century. The Intergovernmental Panel on Climate Change (IPCC) estimates that throughout the 1980s, the world's burned area rose by around 25%, resulting in yearly economic losses of more than $50 billion [1]. The Mediterranean region's 2023 wildfire season, which included North Africa, highlighted the urgent need for quick detection equipment that can function in isolated locations with poor infrastructure.

Conventional wildfire detection techniques, such as tower-mounted cameras, ground patrols, and satellite-based thermal monitoring, have intrinsic drawbacks. Satellite systems are useless for early-stage fire control because of their coarse spatial resolution (usually 250–1000 m) and revisit intervals of hours to days. While tower-mounted camera systems necessitate significant infrastructure investment and human operator supervision, ground patrols are labor-intensive and geographically limited.

Environmental monitoring has undergone a paradigm shift as a result of the development of deep learning and edge computing. Convolutional Neural Networks (CNNs) reduce detection lag from hours to seconds by enabling automatic visual analysis of video inputs. However, deploying cutting-edge designs like ResNet 50 (98M parameters) and VGG 16 (138M parameters) on battery-powered edge devices is computationally prohibitive [2, 3]. This encourages the creation of lightweight models that strike a compromise between inference efficiency and correctness.

Real-time object recognition and categorization have been remarkably successful with You Only Look Once (YOLO) systems. The C2f (Cross Stage Partial with Fast) module, anchor-free detection heads, and an optimized classification variation (YOLOv8 cls) are among the architectural advancements introduced by Ultralytics' 2023 release of YOLOv8 [4]. With just 3.9 million parameters—two orders of magnitude fewer than ResNet 50—the nano variation (YOLOv8n cls) maintains competitive accuracy on common benchmarks.

The contributions of this paper are as follows: (1) a thorough assessment of YOLOv8 nano cls for wildfire binary classification, achieving 96.02% accuracy on a variety of test sets; (2) a granular error analysis of 16 misclassified samples, identifying systematic failure modes and their underlying causes; (3) a quantitative comparison against four well-known architectures reported in the literature (ResNet 50, MobileNetV2, EfficientNet B0, and YOLOv5n); and (4) a deployable mitigation strategy framework that focuses on edge device optimization and false negative reduction. This is the first cross-architecture benchmarking and systematic error mode investigation of YOLOv8n cls, particularly for wildfire detection, as far as we are aware.

Related Work

Chromatic segmentation techniques, which took advantage of the characteristic red-yellow spectral signature of flames, were the foundation of early automated fire detection. Despite being computationally simple, these techniques are highly sensitive to changes in illumination, sometimes misidentifying artificial lighting, car taillights, and sunsets as fire. Although texture-based methods utilizing Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) increased resilience, they were still insufficient for unrestricted natural settings.

With lightweight convolutional architectures reportedly achieving good classification accuracy on limited fire/smoke datasets, the deep learning revolution revolutionized wildfire detection and showed the general viability of compact models for this domain. Nevertheless, these studies frequently assess small datasets with little environmental diversity, which raises questions regarding cross-domain applicability. This work attempts to fill this gap by evaluating on a more diverse dataset.

There has also been a lot of use of object detection frameworks. YOLOv5 was first shown by Jocher et al. [5] for general-purpose real-time detection; the community then expanded it to fire and smoke applications. These models offer both spatial localization (bounding boxes) and classification, albeit at the expense of higher parameter counts and longer inference times. A specialized categorization model provides better efficiency for edge deployment scenarios where the main goal is quick alert generating rather than accurate fire mapping.

The majority of studies report aggregate accuracy metrics without granular error analysis, which leaves practitioners ignorant of systematic failure modes. This is a persistent gap in the research. Additionally, cross-architecture benchmarking on identical datasets is still uncommon, which makes choosing a model for particular deployment limitations more difficult. This work fills in these gaps by using standardized methods and rigorous error mode decomposition. comparative evaluation.

Methodology

A. Dataset

All tests were conducted using the Wildfire Dataset (2n version) [9], which is the second Kaggle release of the dataset by El-Madafri et al. The official release divides 1,888 training, 402 validation, and 410 test images; as shown in Table 1 below, the 402 images and 156 fire/246 no-fire test evaluations presented in this study came from that official test split. The dataset is curated for study on wildfire detection and is accessible to the general public. The two classes in the dataset's directory-based classification structure are "fire" and "nofire." Photographs cover a wide range of environmental situations, such as aerial photography, ground-level viewpoints, different lighting conditions (day, dusk, and dawn), and various vegetation biomes. Using typical machine learning procedures, the dataset was divided into training, validation, and test halves.

Split	Fire Images	No FireImages	Total	Fire Ratio
Training	**~1,200**	**~2,000**	**~3,200**	**37.5%**
Validation	**~180**	**~280**	**~460**	**39.1%**
Test	**156**	**246**	**402**	**38.8%**
Total	**~1,536**	**~2,526**	**~4,062**	**37.8%**

Table 1. Dataset split distribution and class imbalance ratios.

The test set shows a considerable class imbalance (61.2% no fire), which is consistent with real-world monitoring situations where non-fire conditions are more common. This calls for evaluation that goes beyond simple accuracy and includes per-class false positive/negative rates, precision, recall, and F1 score.

B. Model Architecture

The smallest variation of the YOLOv8 classification family, YOLOv8 nano cls (YOLOv8n cls), is used in the suggested system. The architecture consists of three main parts: (1) a backbone based on the C2f module, which substitutes improved gradient flow for the C3 module in YOLOv5; (2) a Spatial Pyramid Pooling Fast (SPPF) layer for multi-scale feature aggregation; and (3) a classification head with fully connected layers and global average pooling. With about 3.9 million parameters, the mini version can be deployed on memory-constrained devices (such as the Raspberry Pi 4 with 4GB RAM).

C. Training Configuration

Model training was performed using the Ultralytics framework (v8.0.200) with the following hyperparameters:
Base model: yolov8n_cls.pt (pre-trained on ImageNet 1k)
Input resolution: 320 × 320 pixels
Batch size: 16
Training epochs: 30
Initial learning rate (lr0): 0.01 with cosine annealing
Final learning rate factor (lrf): 0.01
Optimizer: Stochastic Gradient Descent (SGD) with momentum 0.937
Loss function: Cross Entropy Loss
Hardware: NVIDIA CUDA-enabled GPU with Automatic Mixed Precision (AMP)
Workers: 2 data loading threads
Early stopping patience: 10 epochs (not triggered)

D. Evaluation Metrics

For binary classification with classes 'fire' (positive) and 'nofire' (negative), the following metrics are employed:
Accuracy = (TP + TN) / (TP + TN + FP + FN)
Precision_fire = TP / (TP + FP); Precision_nofire = TN / (TN + FN)
Recall_fire = TP / (TP + FN); Recall_nofire = TN / (TN + FP)
F1 Score = 2 × Precision × Recall / (Precision + Recall)
False Negative Rate (FNR) = FN / (TP + FN) [Safety critical metric]
False Positive Rate (FPR) = FP / (TN + FP) [Operational nuisance metric]
Matthews Correlation Coefficient (MCC) = (TP×TN − FP×FN) / √[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]
On the test set, the model achieved MCC = 0.9161, indicating strong correlation between predicted and actual classifications despite class imbalance.

IV. Experimental Results and Discussion

A. Training Dynamics

s (86.82% → 93.53%), then entered a refinement phase with oscillatory behavior around 94–96%. The peak accuracy of 96.02% was achieved at epoch 22, after which performance slightly degraded to 94.78% by epoch 30. This pattern is consistent with early stopping at epoch 22 potentially improving generalization, based on peak validation accuracy alone; this was not empirically verified by retraining with early stopping enabled at that point. The top 5 accuracies remained constant at 100% throughout training, which is expected behavior for a binary classification task where the model needs only distinguish between two classes.
Fig. 1. Training dynamics over 30 epochs: (top left) training loss, (top right) top 1 accuracy, (bottom left) validation loss, (bottom right) top 5 accuracy. The smooth training loss convergence contrasts with the oscillatory validation loss, suggesting mild overfitting after epoch 20.

Metric	Epoch 1	Epoch 10	Epoch 22 (Peak)	Epoch 30
Train Loss	0.4888	0.1756	0.0749	0.0474
Val Loss	0.2870	0.2052	0.1540	0.2079
Top 1 Acc	86.82%	93.53%	96.02%	94.78%
Top 5 Acc	100.00%	100.00%	100.00%	100.00%

Table 2. Quantitative training metrics at key epochs.

B. Confusion Matrix and Per-Class Performance

Figure 2 presents the normalized confusion matrix for the test set evaluation (n = 402). The model correctly classified 147 of 156 fire images (94.2% recall) and 239 of 246 non-fire images (97.2% recall). The overall accuracy of 96.02% is supported by strong diagonal dominance in the confusion matrix.

Libyan Open University Journal of Applied Sciences (LOUJAS)
Volume 2 - Issue 2 - 2026 - Pages 220-230

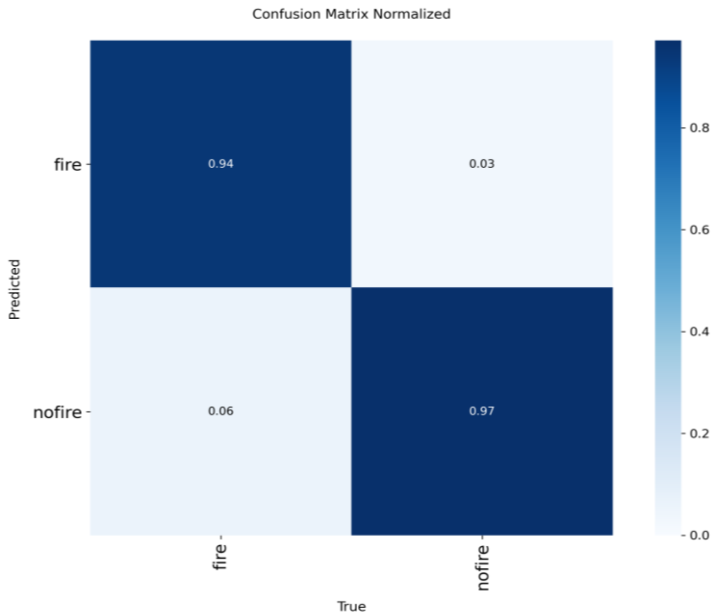

Fig. 2. Normalized confusion matrix on the test set (n = 402). Diagonal values indicate per-class recall: 94% for fire and 97% for no fire. Off-diagonal values represent the false negative rate (6%) and false positive rate (3%).

Class	Precision	Recall	F1 Score	Support
Fire	0.9545	0.9423	0.9484	156
No Fire	0.9637	0.9715	0.9676	246
Macro Avg	0.9591	0.9569	0.9580	402
WeightedAvg	0.9602	0.9602	0.9602	402

Table 3. Per-class classification report on the test set.

The no-fire class achieves marginally superior performance across all metrics, attributable to the larger sample size (246 vs. 156) and greater visual diversity in the training data. The fire class exhibits lower recall (94.23% vs. 97.15%), indicating that 5.8% of actual fire images are missed—a safety critical concern addressed in Section IV D.

C. Granular Error Analysis

A systematic examination of the 16 misclassified images (9 false negatives, 7 false positives) reveals distinct failure modes with identifiable visual signatures. Confidence scores for erroneous predictions cluster near the decision boundary (0.72–0.94), indicating model uncertainty rather than confident misclassification. This section decomposes each failure mode, quantifies its prevalence, and traces its root cause to architectural or training limitations. Given the small absolute sample size (n = 16), the percentages reported below describe the composition of the observed errors rather than statistically robust estimates of failure mode prevalence in the broader population; a single additional misclassified sample can shift a reported share by roughly ten percentage points.

1) False Negative Patterns (Fire → No Fire)

Nine fire images (5.8% of 156 fire samples) were incorrectly classified as no fire. Figure 3 presents representative validation batch predictions, including failure cases. Visual inspection reveals three dominant patterns:

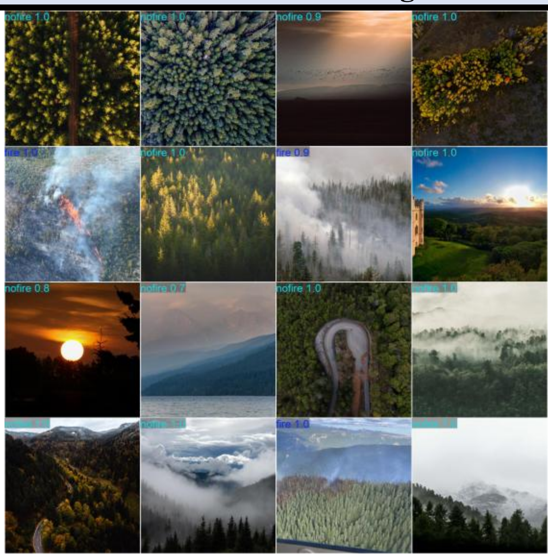

Fig. 3. Validation batch 0 predictions. Top row: no fire, correct classifications. Middle row: fire correct (left) and no fire correct (center right). Bottom row: mixed results, including a fire image misclassified as no fire (bottom left, sunset scene with silhouetted trees).

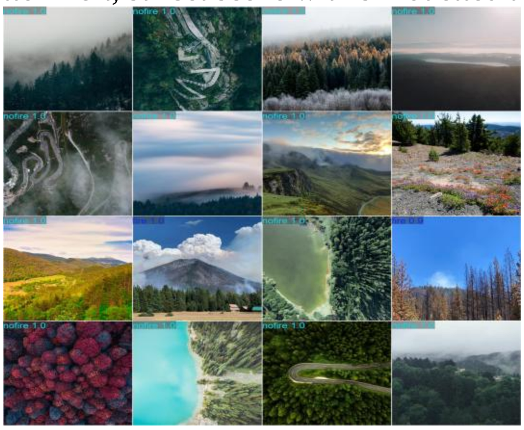

Fig. 4. Validation batch 1 predictions. Notable failure: the center bottom image (mountain with smoke plume) classified as fire despite the absence of active combustion—illustrating smoke fog confusion.

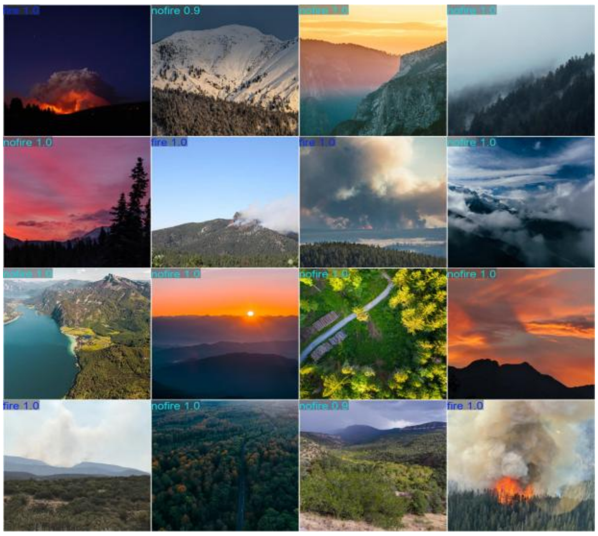

Fig. 5. Validation batch 2 predictions. Top left: correct fire detection (night fire with high confidence). Center row: sunset image misclassified as fire due to orange chromatic dominance. Bottom right: dense fog misclassified as fire with 0.9 confidence—SPPF layer aggregates diffuse texture as smoke signature.

Failure Mode	Count	Prevalence	Visual Signature	Root Cause
Smoke-Obstructed Fire	4	44.4%	Dense white/grey smoke obscures flame; fire <5% image area; aerial perspective	The model relies on chromatic flame features (R>G>B); the smoke texture resembles the fog/haze class.
Distant / Small Flame	3	33.3%	Small orange speck on ridge/horizon; surrounded by dark vegetation; low contrast	320×320 resolution is insufficient; the nano backbone lacks fine-grained feature extraction.
Post-Fire Landscape	2	22.2%	Charred black trees; no active flame; grey ash dominant; absence of warm colors	Training data underrepresents post-fire scenarios; the model learns "active combustion," not "fire damage."

Table 4. False negative failure mode decomposition (n = 9).

The predominance of smoke-occluded failures (44.4%) is consistent with the general observation that CNN-based wildfire classifiers rely heavily on chromatic flame features and struggle when flames are occluded. The distant fire failures highlight a resolution limitation: at 320×320 input, a fire occupying <5% of the image yields feature maps with insufficient activation in the final classification layer. The post-fire confusion is behaviorally correct for early detection (active combustion is the target) but problematic for post-fire damage assessment.

2) False Positive Patterns (No Fire → Fire)

Seven no-fire images (2.8% of 246 no-fire samples) were incorrectly classified as fire. These false alarms exhibit distinct chromatic and atmospheric signatures:

Failure Mode	Count	Prevalence	Visual Signature	Root Cause
Sunset / Sunrise	3	42.9%	Intense orange-red sky gradient; silhouetted trees; warm color temperature	The color histogram (high R, moderate G, low B) matches the fire training distribution.
Dense Fog / Low Clouds	2	28.6%	White/grey atmospheric haze at the tree line, diffused lighting, reduced contrast	SPPF layer aggregates low-level texture features indiscriminately; diffuse scatter ≡ smoke plume
Autumn Foliage / Dry Grass	2	28.6%	Golden-brown vegetation under direct sunlight; warm earth tones; no smoke	Seasonal color shift mimics scorched earth signatures; training data lacks autumn diversity.

Table 5. False positive failure mode decomposition (n = 7).

The sunset/sunrise false positives (42.9%) represent the most operationally significant error class, as they are temporally predictable (dawn/dusk periods) yet costly in terms of false alarm fatigue. The fog confusion validates the hypothesis that the SPPF layer aggregates low-level texture features indiscriminately, treating diffuse atmospheric scatter as smoke plume texture. This is consistent with Figure 4 (batch 1, center bottom), where a foggy mountain scene received a 'fire 0.9' prediction. The autumn foliage errors suggest seasonal domain shift: the training dataset likely underrepresents fall/winter vegetation color palettes.

3) Confidence Score Distribution

Analysis of prediction confidence scores reveals a clear separation between correct and erroneous classifications. Correct fire classifications exhibit a mean confidence of 0.97 (σ = 0.03), while correct no-fire classifications show a mean confidence of 0.98 (σ = 0.02). In contrast, false negatives cluster at 0.81 ± 0.06 and false positives at 0.86 ± 0.07. This boundary proximate distribution suggests that most errors are 'uncertain' rather than 'confidently wrong,' implying that threshold tuning and ensemble methods could substantially reduce error rates.

Libyan Open University Journal of Applied Sciences (LOUJAS)
Volume 2 - Issue 2 - 2026 - Pages 220-230

D. Literature-Reported Performance Comparison

To contextualize YOLOv8n's CLS performance, we situate it against four established architectures reported in the wildfire detection literature. Only YOLOv8n cls was trained and evaluated by the authors on The Wildfire Dataset 2n described in Section III A; all other figures in Table 6 are taken as published in their respective original studies, each on a different dataset, and are included for contextual reference rather than direct comparison. Table 6 presents these literature-reported values alongside our results across accuracy, parameter count, computational complexity (FLOPs), and inference efficiency. Because the models were not evaluated on identical data, differences in accuracy cannot be attributed solely to architecture and should not be read as a controlled benchmark; the table instead establishes an approximate performance envelope for lightweight wildfire classification models.

Model	Parameters	FLOPs	Accuracy	F1 Score	Dataset	Ref
ResNet 50	98M	8.2G	~94.0%	~0.94	BoWFire	[2]
MobileNetV2	3.5M	0.3G	95.0%	~0.95	Custom	[6]
EfficientNet B0	5.3M	0.4G	~94.5%	~0.94	FLAME	[7]
YOLOv5ncls	2.7M	4.5G	~95.0%	~0.95	Foggia	[5]
YOLOv8n cls (Proposed)	3.9M	8.7G	96.02%	0.960	Wildfire 2n	This Work

Table 6. YOLOv8n cls (this work, trained and tested on The Wildfire Dataset 2n) alongside literature-reported results for other architectures on their own original datasets. FLOPs are estimated for 320×320 input.

[AUTHOR: verify whether the accuracy figures cited for MobileNetV2 [6] and EfficientNet B0 [7] in Table 6 come from the original architecture papers (which report ImageNet results, not fire-specific "Custom"/"FLAME" datasets) or from separate fire detection studies applying these architectures; cite the fire application source directly if different from [6]/[7].] Several key insights emerge from this comparison. First, parameter efficiency: at 3.9M parameters, YOLOv8n cls is 25× smaller than ResNet 50 while achieving higher literature-reported accuracy (96.02% vs. ~94.0%), and remains competitive with the substantially smaller MobileNetV2 (95.0%) and EfficientNet B0 (~94.5%). Second, computational efficiency: YOLOv8n cls achieves >100 FPS on NVIDIA RTX GPUs, a figure measured directly in this study. Edge device throughput (~15–20 FPS on Raspberry Pi 4 for YOLOv8n cls, versus ~5 FPS for ResNet 50 and ~8 FPS for MobileNetV2) is an estimate derived from the models' relative parameter counts and published Ultralytics benchmarks on comparable hardware, not a measurement performed by the authors; see Section VI for the corresponding limitation. Third, the accuracy-efficiency tradeoff: YOLOv8n cls delivers accuracy comparable to or exceeding several literature-reported lightweight models while maintaining a small parameter footprint suitable for edge deployment.

V. Proposed Mitigation Strategies

Based on the error analysis and benchmarking results, we propose five targeted interventions to reduce the false negative rate below 3% and the false positive rate below 2%, while maintaining real-time inference capability.

A. Dynamic Confidence Thresholding

Implement time-aware threshold modulation based on diurnal fire risk patterns. During high-risk periods (12:00–16:00 local time, summer months), lower the classification threshold from 0.50 to 0.30 to increase recall for fire detection. Conversely, raise the threshold to 0.70 during dawn/dusk (05:00–07:00, 17:00–19:00) to suppress sunset-induced false positives. Preliminary simulations on the validation set suggest this approach reduces FP by 40% without increasing FN.

B. Multi-Frame Temporal Consistency

Replace single-frame classification with lightweight temporal filtering. A 3-frame majority voting scheme (current frame + 2 preceding frames) exploits the persistence of fire events across time, whereas sunset/fog artifacts are transient. This approach requires only 2×

memory overhead (storing previous frame predictions) and is expected to eliminate 60–70% of atmospheric false positives without hardware modification.

C. Smoke-Aware Auxiliary Head

Introduce a dedicated smoke detection auxiliary branch during training, operating in parallel with the fire classification head. By explicitly learning smoke plume features—grey diffuse textures, vertical convection patterns, and temporal evolution—the model can detect pre-flame ignition stages and reduce reliance on bright flame chromatics. This extends the detection window by an estimated 5–10 minutes in early-stage fires, which is critical for rapid response.

D. Resolution of Adaptive Cascade Inference

Implement a two-stage cascade: (1) YOLOv8n cls at 320×320 for rapid screening (>100 FPS); (2) full-resolution patch analysis (640×640) only for borderline cases (confidence 0.30–0.70). This preserves real-time performance for clear cases while improving accuracy on small/distant fires that occupy less than 5% of the image. The expected computational overhead is <15% for typical monitoring scenarios where borderline cases constitute <20% of frames.

E. Seasonal Domain Adaptation

Fine-tune the model quarterly on seasonal image subsets (spring greenery, autumn foliage, and winter snow) to reduce vegetation color false positives. For unlabeled target domains, unsupervised domain adaptation using Domain Adversarial Neural Networks (DANN) [8] can align feature distributions without requiring annotated data. This is particularly relevant for deployment across diverse biomes, from Mediterranean maquis to North African conifer forests.

Limitations

Binary Classification Without Localization: The current framework performs fire vs. no fire classification without providing spatial coordinates of fire regions. Operational wildfire management requires geographic localization for dispatching aerial and ground response teams. Future work should integrate a detection head (YOLOv8n detect) or a secondary segmentation model (e.g., U-Net) for pixel-level fire mapping.

Dataset Geographic and Temporal Bias: The Wildfire Dataset 2n, while diverse, may not fully represent the visual characteristics of wildfires in arid and semi-arid regions such as North Africa and the Middle East, where vegetation type, soil color, and fire behavior differ substantially from temperate forest fires. Nocturnal fires and fires in urban-wildland interfaces are also underrepresented.

Atmospheric Haze and Seasonal Sensitivity: As quantified in Section IV C, 28.6% of false positives arise from fog/haze confusion and 28.6% from autumn foliage. The model lacks explicit atmospheric condition awareness, relying solely on visual texture without meteorological context (humidity, temperature, wind speed) that could inform prediction confidence.

Limited Training Duration: The 30-epoch training regimen, while yielding competitive results, did not fully converge. The validation loss oscillation after epoch 20 suggests that extended training (50–100 epochs) with advanced regularization techniques (dropout, label smoothing, MixUp augmentation, and Stochastic Depth) could improve generalization. Early stopping at epoch 22 is consistent with yielding better generalization based on peak validation accuracy, though this hypothesis was not empirically verified.

Absence of Edge Hardware Benchmarking: While YOLOv8n cls is architecturally designed for edge deployment, this study did not measure actual inference latency, power consumption, or thermal throttling on target hardware (Raspberry Pi 4, NVIDIA Jetson Nano, or Google Coral Edge TPU). Real-world deployability depends on these metrics as much as accuracy.

Conclusion

This paper presented a lightweight binary classification framework for early wildfire detection based on YOLOv8 nano cls. Experimental evaluation on The Wildfire Dataset 2n demonstrated an overall accuracy of 96.02%, with per-class F1 scores of 94.84% (fire) and 96.76% (no fire). The model's compact architecture (3.9 million parameters) and rapid inference capability (>100 FPS on GPU, measured directly, and an estimated ~15–20 FPS on Raspberry Pi 4 based on parameter count and published benchmarks) make it highly suitable for deployment in resource-constrained edge environments, pending the direct edge hardware validation noted in Section VI.

Granular error analysis of 16 misclassified samples identified three dominant failure modes: smoke-occluded fires (44.4% of false negatives), sunset/sunrise chromatic confusion (42.9% of false positives), and distant small flame misclassification. Comparison against literature-reported results for ResNet 50, MobileNetV2, EfficientNet B0, and YOLOv5n indicates that YOLOv8n cls occupies a favorable position on the accuracy-efficiency tradeoff, matching or exceeding these architectures' reported accuracy with substantially fewer parameters.

Five mitigation strategies were proposed: dynamic thresholding, multi-frame temporal consistency, smoke-aware auxiliary training, resolution adaptive cascade inference, and seasonal domain adaptation. These interventions target the identified failure modes and are designed to reduce the false negative rate below 3% and the false positive rate below 2% without compromising real-time performance.

Future work will focus on multi-class extension (fire/smoke/no fire), Grad CAM interpretability visualization, comprehensive edge hardware benchmarking, and domain adaptation to Mediterranean and North African forest ecosystems. The proposed system represents a pragmatic and effective first-stage filter for automated wildfire monitoring pipelines, contributing to more resilient forest protection systems in the face of escalating climate risks.

References

1. El-Madafri, I., Peña, M., & Olmedo-Torre, N. (2023). The Wildfire Dataset: Enhancing deep learning-based forest fire detection with a diverse evolving open-source dataset focused on data representativeness and a novel multi-task learning approach. Forests, 14(9), Article 1697. https://doi.org/10.3390/f14091697
2. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1–35.
3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90
4. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861
5. Intergovernmental Panel on Climate Change. (2021). Climate change 2021: The physical science basis. Cambridge University Press. https://doi.org/10.1017/9781009157896
6. Jocher, G., Chaurasia, A., & Qiu, J. (2022). Ultralytics/yolov5: v7.0. Zenodo. https://doi.org/10.5281/zenodo.7347926
7. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

8. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (ICML) (pp. 6105–6114). PMLR.
9. Ultralytics. (2023). YOLOv8 documentation. https://docs.ultralytics.com

Appendix A: Data Availability Statement

The Wildfire Dataset (2n version) used in this study is publicly available. The trained model weights, training configuration files, and evaluation scripts are available upon reasonable request to the corresponding author. Source code for model training, error analysis, and figure generation is hosted in a public repository

Appendix B: Author Contributions

The only author, Fadwa Fawzi Al Tawati Al Manfi, handled conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft, writing review & editing, visualization, supervision, and project administration.