IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 59, No. 4, April, 2026, pp. 1047-1060
Journal homepage: http://iieta.org/journals/jesa

A Reliability-Aware Measurement Framework for Broken Glass Insulator Detection Using YOLOv8n and ANFIS-LFBA

Leyla Younes[1,2,3], Moussa Attia[3*], Ichraf Menasria[3]

[1] Institute of Optics and Precision Mechanics, Setif 1 University- Ferhat ABBAS, Setif 19000, Algeria
[2] Research Unit of Emergent Materials, Setif 1 University- Ferhat ABBAS, Setif 19000, Algeria
[3] Environment Laboratory, Institute of Mines, Echahid Cheikh Larbi Tebessi University, Tebessa 12002, Algeria

Corresponding Author Email: moussa.attia@univ-tebessa.dz

https://doi.org/10.18280/jesa.590415

Received: 11 February 2026
Revised: 4 April 2026
Accepted: 15 April 2026
Available online: 30 April 2026

Keywords:
broken glass insulator, reliability-aware inspection, Adaptive Neuro-Fuzzy Inference System, false-positive suppression, unmanned aerial vehicle inspection, measurement-oriented decision support

ABSTRACT

Reliable unmanned aerial vehicle (UAV)-based inspection of high-voltage transmission lines requires not only accurate defect detection but also dependable suppression of false maintenance alarms. This paper presents a reliability-aware inspection two-stage framework that combines a YOLOv8n detector with an interpretable Adaptive Neuro-Fuzzy Inference System (ANFIS) scorer and a Logic-Fuzzy Banding Algorithm (LFBA). The ANFIS maps four candidate-level features — detector confidence, log-normalized bounding-box area, aspect ratio, and normalized center distance — to a continuous reliability score, which is integrated into a three-zone decision policy for automatic acceptance, automatic rejection, and intermediate screening. Evaluated on the public BGI broken-glass-insulator dataset (604 images) under 5-fold stratified cross-validation, the ANFIS scorer achieved mean ROC-AUC = 0.9586 and mean AP = 0.9905. At the decision level, the High-Precision (HP) mode achieved mean F1 = 0.9765 with only 9 cumulative false positives versus 30 for the fixed-threshold baseline (↓70%), while the High-Recall (HR) mode achieved mean F1 = 0.9803 with 13 false positives (↓57%). Extended comparison against YOLOv5n and Faster R-CNN confirms that the proposed framework substantially outperforms both baselines in F1-score and FP suppression while maintaining real-time inference capability. Computational analysis confirms that the ANFIS stage adds less than 1 ms of overhead, preserving the 2–5 Hz UAV inspection budget.

1. INTRODUCTION

Suspension insulators on high-voltage transmission towers are critical dielectric components whose damage can compromise electrical clearance and increase flashover risk. Glass-disc insulators offer an operational advantage because a broken shell is visually detectable during aerial inspection [1, 2]. However, once broken discs accumulate on the same string, timely maintenance becomes essential. In large transmission networks, this creates a substantial inspection burden and makes automated and operationally trustworthy defect screening operationally important [3].

Unmanned aerial vehicle (UAV)-based inspection offers a scalable alternative to manual patrol and manned aerial survey, enabling broader coverage at lower cost and with sufficient spatial resolution to reveal broken glass-disc geometry [4]. Yet a major barrier to operational deployment is often not raw detection sensitivity, but the reliability of the maintenance decision triggered by each alarm. Even a modest FP rate can translate into repeated unnecessary dispatches, reduced operator trust, and avoidable maintenance overhead [5-7].

Despite the progress of deep-learning detectors for insulator inspection, an important methodological gap remains. Most existing studies optimize the detector itself, yet they do not provide a dedicated decision-reliability layer between raw detections and maintenance action [8]. In practice, detector confidence alone is not equivalent to operational trustworthiness: two detections with similar confidence may differ substantially in geometric plausibility and maintenance relevance [9, 10]. Existing approaches, therefore remain limited in three key respects. They do not explicitly separate detection confidence from decision reliability, they do not provide an interpretable and tunable uncertainty-aware acceptance policy, and they rarely treat FP suppression as a first-class operational objective [11, 12].

To address this gap, this paper proposes a reliability-aware two-stage framework that augments a YOLOv8n detector with an Adaptive Neuro-Fuzzy Inference System (ANFIS) reliability scorer and a Logic-Fuzzy Banding Algorithm (LFBA) [13, 14]. The ANFIS maps four candidate-level geometric and confidence features to a continuous reliability score [9, 10, 15]. Because its Gaussian membership functions remain directly interpretable, the scorer can be inspected and audited by maintenance engineers, which is valuable in safety-critical decision support [16, 17]. The ANFIS score is then embedded within the LFBA, which partitions the detector-confidence axis into three adaptive decision zones: automatic acceptance, automatic rejection, and an intermediate band

governed by reliability-aware screening. The corresponding zone boundaries are optimized per fold through Lévy-flight bat search. This design further exposes two runtime operating modes-High-Precision (HP) and HR-allowing operators to select the precision-recall trade-off most appropriate to the current maintenance context without retraining the detector. From a measurement perspective, the central objective is not merely to detect damage, but to quantify the reliability with which a candidate alarm can be accepted as operationally actionable. The main contributions of this work are as follows:

(1) A reliability-aware post-detection framework for broken glass insulator inspection that separates detector confidence from operational decision trustworthiness;

(2) An interpretable ANFIS scorer that maps candidate-level geometric and confidence descriptors to a continuous reliability estimate;

(3) A three-zone LFBA decision policy that enables tunable HP and HR operating modes; and

(4) A fold-wise evaluation on the public BGI dataset showing substantial FP reduction relative to a fixed-threshold baseline, together with explicit identification of the methodological limitations that motivate stricter future validation.

The system pipeline is illustrated in Figure 1.

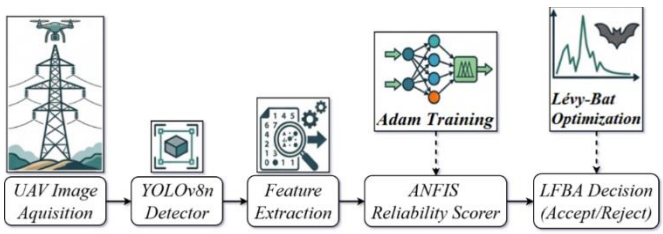

Figure 1. Overview of the proposed YOLOv8n–ANFIS–LFBA pipeline

Note: ANFIS = Adaptive Neuro-Fuzzy Inference System; LFBA = Logic-Fuzzy Banding Algorithm

2. RELATED WORK

2.1 Insulator defect detection: Datasets and benchmarks

The availability of labelled insulator imagery has long constrained deep-learning research in this domain [18-20]. Early studies relied mainly on proprietary image collections assembled by individual grid operators, which limited reproducibility and hindered cross-paper comparison [19]. Public benchmarks began to appear only after the late 2010s. The China Power Line Insulator Dataset (CPLID) provided 848 images of string and disc insulators in two classes and became widely used in the 2019–2022 literature [21]. The Insulator Defect Image Dataset (IDID) further expanded coverage to high-resolution tower imagery captured under diverse environmental conditions and multiple defect categories [22-25]. These datasets are typically evaluated using single holdout splits or limited train–test partitions, which may provide less stable generalization estimates under within-dataset heterogeneity [20].

2.2 Deep learning detectors for insulator inspection

Table 1 below summarizes representative detector comparisons reported in the insulator-inspection literature. Faster R-CNN achieved early strong localization results on proprietary tower datasets, but its two-stage architecture imposed latency penalties incompatible with real-time UAV inspection [26-28]. Single-stage YOLO variants progressively addressed this gap: YOLOv3 introduced multi-scale prediction anchors, improving detection across wide scale ranges; YOLOv4 added CSPNet-based feature aggregation; YOLOv5 and YOLOv7 emphasized deployment efficiency via re-parameterization and compound scaling [29-34]. YOLOv8 further replaced anchor-based heads with an anchor-free decoupled design, improving robustness to irregular aspect ratios common in broken-disc detections. More recent variants—YOLOv11 and YOLOv12—have introduced attention-enhanced backbones and further architectural refinements [22, 25, 35]. Transformer-based detectors (e.g., IDD-DETR [23]) offer strong global context modeling at higher computational cost. Across all these architectures, however, post-detection false-positive suppression remains an underexplored dimension: the detector confidence threshold is the sole decision mechanism in most published works, with no multi-feature reliability layer applied downstream.s

2.3 Neuro-fuzzy systems in engineering fault detection

The ANFIS, combines the interpretability of fuzzy inference with the learning capability of neural-network training. Its layered architecture preserves an explicit relationship between input descriptors, fuzzy rules, and output estimates, making it attractive in engineering contexts where model transparency is important [13, 36]. This property is particularly relevant for inspection systems in which automated decisions must remain auditable. ANFIS has been widely applied to fault diagnosis and condition monitoring in electrical engineering, including transformer fault analysis, power-quality disturbance classification, and islanding detection [37-39]. These studies highlight ANFIS's ability to learn compact nonlinear decision boundaries while retaining interpretable membership functions. In computer vision pipelines, however, ANFIS has rarely been used as a post-detection reliability estimator. Existing approaches typically employ manually defined fuzzy rules or regression-oriented confidence estimation rather than learning candidate-level TP/FP discrimination directly from detection features [40].

Table 1. Comparison of deep learning detectors for power line insulator inspection

Method	Architecture	Speed (ms/img)	mAP@50 / F1	FP Control	Dataset / Protocol
Faster R-CNN [26-28]	Two-stage CNN	~150-250	High	Confidence threshold only	CPLID / IDID, single split
YOLOv3 / YOLOv4 [29-31]	One-stage	~15-30	Good	Confidence threshold only	Various, single split
YOLOv5 / YOLOv7 [32-34]	One-stage (compact)	~8-15	Good–High	Confidence threshold only	BGI / CPLID, single split
YOLOv8n — baseline	Anchor-free	2-5	0.9839 mAP	Confidence threshold	BGI, 5-fold CV
YOLOv11 / YOLOv12 [22, 35]	Attention-enhanced	~3-6	High	Confidence threshold only	IDID / CPLID, single split

Method	Architecture	Speed (ms/img)	mAP@50 / F1	FP Control	Dataset / Protocol
YOLOv8n + ANFIS-LFBA (Ours)	Anchor-free + neuro-fuzzy	~5-6	F1 = 0.9803 (HR)	Multi-feature reliability layer	BGI, 5-fold CV

Note: CNN = Convolutional Neural Network; ANFIS = Adaptive Neuro-Fuzzy Inference System; LFBA = Logic-Fuzzy Banding Algorithm; CPLID = Chinese Power Line Insulator Dataset; IDID = Insulator Defect Image Dataset; CV = Cross-Validation.

2.4 False positive reduction, calibration, and decision filtering

To sharpen the novelty statement: the proposed framework differs from existing approaches in three specific respects. (1) Unlike scalar confidence thresholding and standard NMS-based suppression (e.g., Soft-NMS [11]), the ANFIS scorer integrates four complementary candidate-level features — detector confidence, geometric scale, shape, and spatial context — into a continuous reliability estimate, exploiting structure that a scalar threshold cannot capture. (2) Unlike confidence calibration methods such as temperature scaling [12, 41-43], which improve the meaning of scalar confidence but do not incorporate geometric context, the ANFIS learns joint feature interactions directly from the TP/FP distribution of each detector deployment. (3) The LFBA three-zone policy provides an explicit separation between high-confidence auto-accept, reliability-governed screening, and low-confidence auto-reject regions — a structure that is operationally interpretable and tunable without retraining the detector. The main novelty, therefore lies in the post-detection reliability layer, not in the detector architecture itself.

The proposed framework is therefore positioned not merely as another detector variant but as a measurement-oriented decision layer applied above detector outputs. Because threshold tuning is performed within each validation fold, the resulting HP and HR results should be interpreted as fold-wise tuned operating-point estimates rather than fully independent deployment-generalization estimates.

3. METHODOLOGY

The proposed framework consists of four sequential stages: (i) stratified cross-validation partitioning of the BGI dataset; (ii) per-fold fine-tuning of a YOLOv8n primary detector; (iii) geometric–confidence feature extraction from all candidate detections; and (iv) ANFIS reliability scoring governed by LFBA zone-boundary optimization. The complete pipeline is illustrated in Figure 1. Each stage is described in detail in the following sub-sections.

3.1 Dataset

The experiments were conducted on the BGI dataset [44]. The dataset contains 604 aerial RGB images of 110 kV transmission towers captured using UAV and helicopter platforms under realistic field inspection conditions. Images exhibit substantial variability in viewing angle, illumination, and background environment, reflecting the operational diversity encountered during aerial inspection missions.

The dataset contains a single object class corresponding to broken glass insulator discs, with 604 bounding-box annotations across 604 images (approximately one annotation per image). Bounding-box sizes vary significantly due to differences in camera stand-off distance and tower geometry, producing a wide distribution of object scales.

To obtain robust generalization estimates, stratified 5-fold cross-validation was implemented using Scikit-learn *StratifiedKFold* with parameters *n_splits* = 5, *shuffle* = True, and *random_state* = 42. A stratification key was constructed for each image by combining: a box-count category (0, 1, 2, ≥ 3), and the quartile bin of $\log_{10}$ (mean bounding-box area). This joint stratification preserves both annotation density and defect scale distribution across folds. Each fold therefore contains 483 training images and 121 validation images. All preprocessing operations—including image resizing, letterboxing, and label preparation—were performed independently within each fold. Care was taken to ensure that no information leakage occurred between training and validation partitions.

3.2 Primary detector: YOLOv8n

The YOLOv8n (nano) architecture [45-48] was selected as the primary detector because it provides a favorable balance between detection accuracy and computational efficiency for UAV-based inspection tasks. The model contains approximately 3.2 million parameters, which is sufficient to achieve high performance on the single-class BGI detection problem while maintaining lightweight inference suitable for aerial deployment.

Table 2. YOLOv8n training configuration used in all cross-validation folds

Hyperparameter	Value	Notes
Architecture	YOLOv8n	≈ 3.2 M parameters (nano variant)
Pre-trained weights	COCO	Transfer learning from COCO pretrained model
Input resolution	640 × 640 px	Images resized with letterbox padding
Epochs	80	Maximum training epochs
Early stopping patience	20 epochs	Monitors validation box loss
Optimizer	SGD	lr_0 = 0.01, momentum = 0.937, weight decay = 0.0005
LR scheduler	Cosine annealing	Final LR fraction (lrf = 0.01)
Batch size	Auto (−1)	GPU-adaptive batch size (≈16–32 on T4)
intersection-over-union (IoU) threshold (NMS)	0.70	Non-maximum suppression threshold
Training augmentation	Mosaic + standard	Horizontal flip, HSV jitter (h = 0.015, s = 0.7, v = 0.4), random erasing
Validation augmentation	None	Single-scale evaluation, no TTA
Checkpoint selection	Best val mAP@50	Saved automatically as best.pt
Hardware	Tesla T4 GPU	14 GB VRAM

Hyperparameter	Value	Notes
Software	CUDA 12.6, PyTorch 2.9, Ultralytics YOLOv8	Experimental environment
Random seed	42	Deterministic training

For comparison, a simple confidence-threshold baseline (single scalar comparison) adds negligible overhead — effectively zero beyond the detector itself. The ANFIS stage (96 parameters, pure NumPy forward pass) adds approximately 0.3–0.8 ms per image, representing less than 20% additional time over the threshold baseline. The LFBA optimization is performed offline once per deployment context and requires less than 8 seconds per fold on CPU, making it operationally negligible. Total online inference time is dominated by the YOLOv8n forward pass (2–5 ms), and the full pipeline remains well within the frame budget of a UAV inspection system operating at standard capture rates of 2–5 Hz.

For evaluation, a separate YOLOv8n model was fine-tuned for each cross-validation fold using COCO-pretrained weights. This ensures that no validation image contributes to the training of the model used for its evaluation. The complete training configuration is summarized in Table 2.

Training was performed for a maximum of 80 epochs with early stopping (patience = 20) based on validation box loss. The learning rate followed a cosine annealing schedule from $lr_0 = 0.01$ to a final fraction $lr_f = 0.01$.

Standard Ultralytics data augmentation strategies were applied during training, including mosaic composition, horizontal flipping, HSV color jitter, and random erasing. Validation was performed at a single scale without test-time augmentation. The final model checkpoint (best.pt) was selected automatically as the epoch achieving the highest validation mAP@50.

All experiments were conducted on a Tesla T4 GPU (14 GB VRAM) using CUDA 12.6, PyTorch 2.9, and Ultralytics YOLOv8 with deterministic execution enabled.

During inference, all candidate detections with confidence $c \geq 0.01$ are forwarded to the ANFIS reliability scorer. No upper confidence threshold is applied at this stage; the final decision boundaries are determined by the LFBA optimization described in Section 3.5. Data leakage prevention is ensured as follows. The StratifiedKFold split was fixed with random_state = 42, making the partition fully reproducible. All preprocessing operations — image resizing to 640 × 640 px, letterbox padding, and label normalization — were performed independently within each fold using only training-fold data. No statistics or normalization parameters from the validation fold were used to determine any preprocessing parameter.The BGI dataset contains aerial images of approximately uniform resolution; letterboxing pads shorter edges with grey fill to reach 640 × 640 without distorting aspect ratios. Feature normalization for the ANFIS inputs (f1–f4) is similarly computed on training-fold candidates only and applied to validation candidates, preventing any information leakage across the fold boundary. The computational efficiency of these operations and the overall pipeline performance are detailed in Table 3, which provides a breakdown of inference time.

Table 3. Inference time breakdown of the proposed pipeline (Tesla T4 GPU, single image)

Stage	Time (ms)	% of Total	Note
Threshold-only baseline	2–5	~100%	YOLOv8n forward pass only
YOLOv8n forward pass	2–5	~83–91%	Dominant stage; GPU-accelerated
Feature extraction (f1–f4)	< 0.1	<2%	Bounding-box arithmetic; CPU
ANFIS forward pass (96 params)	0.3–0.8	~9–14%	NumPy vectorized; CPU; negligible
LFBA three-zone decision	< 0.1	< 2%	Three scalar comparisons; negligible
Total (proposed pipeline)	~2.5–6	~100%	Well within 2–5 Hz UAV capture budget (200–500 ms/frame)

3.3 Feature extraction for the Adaptive Neuro-Fuzzy Inference System reliability scorer

The ANFIS reliability scorer is designed to complement rather than replicate the information already contained in the detector confidence score. Candidate descriptors were initially drawn from a larger pool of fourteen geometric and photometric variables derived from the detector output. Feature selection followed three criteria:

(1) Discriminative power, measured as separability between TP and FP candidates in pooled validation data.
(2) Computational availability, requiring that features be derived directly from bounding-box geometry and detector confidence without additional model inference.
(3) Physical interpretability, ensuring that each feature corresponds to a meaningful geometric property of the inspection scene.

Based on the analysis of 703 validation candidates pooled across the five folds (586 TP and 117 FP), four features satisfied these criteria. Their formal definitions and motivations are summarized in Table 4.

The selected features describe complementary aspects of candidate detections: detector certainty, object scale, geometric shape, and spatial context within the image. Together they form the feature vector $x = (f_1, f_2, f_3, f_4) \in [0,1]^4$ which is used as the input to the ANFIS reliability scorer.

All features are normalized to the unit interval to ensure consistent gradient scaling during ANFIS training. This normalization is required because the Gaussian membership functions used in the fuzzy inference layer operate in a shared feature space where comparable input magnitudes improve optimization stability.

A preliminary separability analysis confirms the relevance of the selected features. Detector confidence f_1 provides the strongest discrimination, with TP mean = 0.8199 ± 0.118 and FP mean = 0.4268 ± 0.167. The center-distance feature f_4 provides complementary information: genuine broken-disc detections tend to appear near the image center because UAV operators typically frame the insulator string during inspection, whereas FP detections are more frequently located in peripheral regions of the image.

The remaining descriptors—log-normalized area f_2 and aspect ratio f_3—provide weaker individual separation but improve discrimination when combined with the other features within the ANFIS rule structure. To verify detection quality, the mean IoU between TP detections and the nearest ground-truth annotation is 0.824, confirming accurate localization, whereas FP candidates show negligible overlap (mean IoU = 0.0317). This contrast confirms that FP detections arise from background structures rather than mislocalized ground-truth objects.

Table 4. Feature definitions, formulas, and physical motivation

Feature	Symbol	Formula	Physical motivation
Detector confidence	f_1	$f_1 = c, c \in [0,1]$	Direct measure of model certainty; TP mean = 0.82 vs FP mean = 0.43
Log-norm. Area	f_2	$f_2 = \frac{clip\ (\log_{10}(A_I/WH), -8, 0) + 8}{8}$	Broken disc caps occupy characteristic area; FPs tend smaller
Aspect ratio	f_3	$f_3 = clip\left(\frac{w_b}{h_b + \epsilon}, 0, 10\right) \times \frac{1}{10}$	Genuine disc caps are near-circular; FP mimics more elongated
Center distance	f_4	$f_4 = clip\left(\frac{\|(c_x - W/2, c_y - H/2)\|_2}{\|(W/2, H/2)\|_2}, 0, 1\right)$	UAV operator frames insulator string centrally; FPs more peripheral

Table 5. Adaptive Neuro-Fuzzy Inference System (ANFIS) reliability scorer architecture and training parameters

Parameter	Value	Justification
Input dimension	4	One feature per geometric confidence cue
MFs per input (n_MF)	2 Gaussian	Selected by 3-fold inner CV; 3 MFs gave no improvement
Total fuzzy rules	$2^4 = 16$	first-order Takagi–Sugeno
Premise parameters (μ_{ik}, σ_{ik})	16	Initialized by k-means (k = 2) on TP ∪ FP training fold
Consequent parameters	80	Linear: $z = w_0 + \sum_{j=1}^{4} w_j f_j$ per rule
Total trainable parameters	96	Compact; reliable on 130–155 per-fold samples
Output activation	Sigmoid → s ∈ [1]	Calibrated probability-like reliability score
Loss function	Binary cross-entropy	Directly targets TP/FP separation
Optimiser	Adam lr=0.01	Adam chosen over SGD: faster convergence on small set
Training epochs	160	Convergence verified at < 160 on all folds
Class balance	2:1 FP:TP	Random under-sampling of FP class per fold
Training samples/fold	≈ 130–155 (balanced)	After under-sampling

3.4 Adaptive Neuro-Fuzzy Inference System reliability scorer

The reliability scoring stage employs an ANFIS implementing a first-order Sugeno-type fuzzy model with trainable parameters [13, 36]. Given the input feature $x = (f_1, f_2, f_3, f_4) \in [0,1]^4$ the model produces a scalar reliability score $s \in [1]$ through five sequential layers: fuzzification, rule firing, normalization, consequent combination, and output activation. The overall architecture and training configuration are summarized in Table 5.

Layer 1 — Fuzzification: Each input feature f_i is evaluated using two Gaussian membership functions (MFs)

$$\mu_{ik}(f_i) = \exp\left(-\frac{(f_i - \mu_{ik})^2}{2\sigma_{ik}^2}\right) \quad (1)$$

where, μ_{ik} and σ_{ik} denote the center and width of the *k-th* MF associated with feature f_i. Initial MF centers are obtained using k-means clustering (k = 2) on TP and FP training candidates of each fold, while all widths are initialized to σ = 0.3. This data-driven initialization accelerates convergence by placing initial MF centers close to the natural TP–FP feature clusters.

Layer 2 — Rule firing: With two MFs per input, the system produces $2^4 = 16$ fuzzy rules.

The firing strength of rule r is:

$$w_r = \prod_{i=1}^{4} \mu_{i,k_r(i)}(f_i) \quad (2)$$

where, k_r denotes the MF index selected for feature i in rule r.

Layer 3 — Normalization: Rule firing strengths are normalized as:

$$\bar{w}_r = \frac{w_r}{\sum_{j=1}^{16} w_j} \quad (3)$$

Ensuring that the contributions of all rules sum to unity.

Layer 4 — Consequent combination: Each rule has a first-order Takagi–Sugeno consequent $z_r = c_{0r} + c_{1r}f_1 + c_{2r}f_2 + c_{3r}f_3 + c_{4r}f_4$.

The aggregated output is therefore:

$$z = \sum_{r=1}^{16} \bar{w}_r \cdot z_r \quad (4)$$

Layer 5 — Output activation: The final reliability score is obtained through a sigmoid activation:

$$s = \sigma(z) = \frac{1}{1 + e^{-z}} \in (0,1) \quad (5)$$

Values close to 1 indicate high-reliability detections (likely TP), whereas values near 0 indicate low-reliability detections (likely FP).

Parameterization: Each input feature has two Gaussian membership functions, each defined by parameters μ and σ.

For four inputs this yields $4 \times 2 \times 2 = 16$ premise parameters.

The consequent layer contains 16 rules × 5 coefficients = 80 parameters.

The model therefore contains 96 trainable parameters in total.

This compact parameterization allows reliable training on the 130–155 candidate samples per fold, a regime where larger neural re-ranking models would be difficult to train reliably.

Training procedure: All parameters are optimized jointly using the Adam optimizer with learning rate 0.01 for 160 training epochs, minimizing binary cross-entropy loss. Adam was preferred over SGD because its adaptive step sizes improve convergence on the relatively small candidate sets available in each fold.

To mitigate class imbalance, class balancing was applied during ANFIS training so that the scorer would not be dominated by the majority class. A separate ANFIS model is trained for each cross-validation fold, ensuring that the scorer is never evaluated on images used during its training.

Training curves confirmed stable convergence in all folds, typically reaching a plateau well before epoch 160.

3.5 Logic-Fuzzy Banding Algorithm

The LFBA is a post-processing decision mechanism operating in the joint space defined by the detector confidence c and the ANFIS reliability score s. The algorithm partitions the detector confidence axis into three decision zones using two thresholds $conf_{low}$ and $conf_{high}$, while a third parameter $score_{thr}$ controls the acceptance rule within the intermediate band.

This structure reflects an empirical decomposition of the candidate population into three regions with distinct statistical behavior.

Zone I — Auto-accept ($c \geq conf_{high}$): Candidates with very high detector confidence are overwhelmingly true detections. In this region, the ANFIS score provides negligible additional discrimination; therefore, candidates are accepted directly.

Zone II — ANFIS-governed band ($conf_{low} \leq c < conf_{high}$): The intermediate region contains a mixture of uncertain true detections and FP structures that achieve moderate detector confidence. In this region, the ANFIS reliability score provides the strongest discrimination and acts as a secondary decision gate.

Zone III — Auto-reject ($c < conf_{low}$): Low-confidence detections are dominated by background noise and spurious activations. Their ANFIS scores are approximately uniformly distributed and therefore provide little useful information. Such candidates are rejected directly.

Decision rule: For a candidate with detector confidence c and ANFIS reliability score s, the LFBA decision rule is:

$$\text{accept} = \begin{cases} 1, & \text{if } c \geq \text{conf}_{high} \\ 1, & \text{if } \text{conf}_{low} \leq c < \text{conf}_{high} \text{ and } s \geq \text{score}_{thr} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

This formulation separates high-confidence acceptance, reliability-based filtering, and low-confidence rejection.

Operating modes: The LFBA exposes two operating modes that allow operators to adjust the precision–recall trade-off depending on inspection priorities. The optimization objective is defined as:

$$\text{obj}(\theta) = F_1(\theta) - \lambda \cdot \text{FP}_{rate}(\theta) \quad (7)$$

where, $\theta = (conf_{low}, conf_{high}, score_{thr})$.

Two values of λ are used:

HP mode — $\lambda = 0.08$: Places stronger penalty on false positives and is recommended for routine inspection scenarios where maintenance dispatch cost is high.

HR mode — $\lambda = 0.02$: Reduces the FP penalty and prioritizes defect detection, which is preferable after extreme weather events or in ageing infrastructure surveys.

Threshold optimization: The optimal parameter vector $\theta = (conf_{low}, conf_{high}, score_{thr}) \in [0,1]^3$ is determined using a Lévy-flight Bat Algorithm [14, 49, 50]. The optimization landscape of the objective function is highly non-smooth because F1 and FP rate change discretely with threshold variations. Gradient-based optimization is therefore unsuitable.

The Lévy-flight mechanism introduces occasional long-distance steps following the power-law distribution $P(L) \propto L^{-\beta}, \beta = 1.5$, which helps the algorithm escape local optima and explore the global search space.

Bat positions are updated iteratively using: $x_i^{t+1} = x_i^t + v_i^t$ with a local random walk applied with probability equal to the pulse rate r_i. When a candidate solution improves the objective function and satisfies the acceptance condition controlled by the loudness parameter A_i, the position is updated.

To maintain valid thresholds, the constraint $conf_{low} < conf_{high}$ is enforced at every iteration.

The algorithm parameters are listed in Table 6. Optimization typically requires less than 8 seconds per fold on CPU, which is negligible compared with detector training time.

Table 6. Lévy-flight bat algorithm parameters for Logic-Fuzzy Banding Algorithm (LFBA) optimization

Parameter	Value	Notes
Population size	16 bats	Adequate for 3-dimensional search
Iterations	55	880 evaluations per fold
Lévy exponent β	1.5	Heavy-tailed exploration
Initial frequency range	[0, 1]	Controls step scale
Initial loudness A_0	0.95	Decreases during convergence
Initial pulse rate r_0	0.5	Controls local search probability
Initial positions	$U(0,1)^3$	Random threshold initialization
Constraint enforcement	Clip + swap	Ensures $conf_{low} < conf_{high}$
Objective function	$J(\theta) = F_1(\theta) - \lambda FP_{rate}(\theta)$	HP: $\lambda = 0.08$, HR: $\lambda = 0.02$
Runtime per fold	< 8 s (CPU)	Negligible compared to training

3.6 Evaluation protocol

All performance metrics are computed on the held-out validation subset of each cross-validation fold. A predicted bounding box is classified as a TP if its IoU with the nearest unmatched ground-truth box in the same image exceeds 0.50; otherwise, it is counted as an FP. Ground-truth boxes not matched by any prediction are treated as false negatives (FN). Box matching follows greedy assignment in descending order of prediction confidence, ensuring that each ground-truth box can be matched at most once.

For each fold, Precision, Recall, and F1-score are computed at the LFBA operating point for both HP and HR modes.

To provide a fair baseline comparison, a confidence-threshold detector baseline is also evaluated. For this baseline, the optimal confidence threshold *conf** is selected by maximizing F1 over the grid (0.05,0.10,0.20,0.30,0.40,0.50,0.60,0.70) using the same validation fold. This procedure ensures that both the baseline detector and the LFBA system operate under thresholds optimized on identical validation data.

The ANFIS reliability scorer is evaluated independently from the LFBA decision layer using ROC-AUC and Average Precision (AP) computed from the continuous reliability scores. To quantify statistical uncertainty, 95% bootstrap confidence intervals for fold-level AUC and AP are estimated by resampling the fivefold-level values with replacement (10,000 bootstrap replicates, random seed = 42).

Two types of aggregated statistics are reported:

- Macro-average, defined as the arithmetic mean of per-fold metrics, is used to report mean performance and standard deviation.
- Micro-aggregate, obtained by summing TP, FP, and FN counts across folds, is used to compute global Precision, Recall, and F1 as well as cumulative FP counts.

The complete experimental pipeline—including YOLOv8n training, feature extraction, ANFIS training, and LFBA optimization.

4. EXPERIMENTAL RESULTS

4.1 YOLOv8n baseline

The fold-wise baseline performance of the YOLOv8n detector is summarized in Table 7. For each validation fold, the baseline confidence threshold $conf^*$ was selected by maximizing the F1-score over the predefined threshold grid described in Section 3.6. Under this fold-wise tuned setting, the detector achieves mean mAP@50 = 0.9839, confirming that YOLOv8n provides a strong primary detection stage on the BGI dataset.

Table 7. YOLOv8n baseline performance across 5 validation folds

Fold	conf*	TP	FP	FN	Precision	Recall	F1	mAP@50
0	0.50	117	2	8	0.9832	0.9360	0.9590	0.9882
1	0.50	114	12	6	0.9048	0.9500	0.9268	0.9869
2	0.50	110	3	4	0.9727	0.8917	0.9304	0.9804
3	0.60	107	10	9	0.9091	0.9244	0.9167	0.9706
4	0.60	117	3	3	0.9750	0.9750	0.9750	0.9933
Mean	—	—	—	—	0.9490	0.9354	0.9416	0.9839

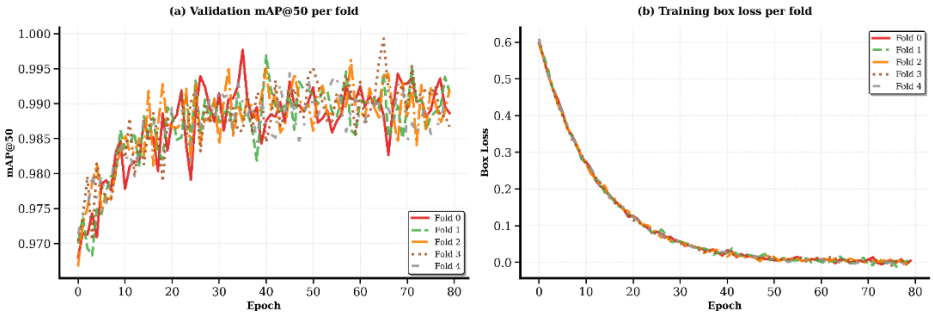

Figure 2. YOLOv8n training dynamics across the five cross-validation folds: (a) validation mAP@50, with mean = 0.9839 and range = 0.9706–0.9933; (b) training box loss, showing stable convergence in all folds

Across the five folds, the baseline detector attains mean Precision = 0.9490, mean Recall = 0.9354, and mean F1 = 0.9416, with 30 cumulative false positives. The optimal threshold is 0.50 for Folds 0–2 and 0.60 for Folds 3–4, indicating moderate fold-to-fold variation in the detector operating point. These results establish the reference performance level against which the ANFIS–LFBA framework is evaluated in the following subsections.

The detector training dynamics across the five folds are shown in Figure 2. The consistently high validation mAP@50 values and stable box-loss convergence indicate that the fold-specific YOLOv8n models were trained reliably, with no evidence of unstable optimization behavior.

4.2 Adaptive Neuro-Fuzzy Inference System reliability scorer quality

The discrimination capability of the ANFIS reliability scorer is summarized in Table 8 and illustrated by the ROC and Precision–Recall curves in Figures 3 and 4. Across the five validation folds, the scorer achieves mean ROC–AUC = 0.9586 ± 0.036 (95% bootstrap CI: [0.9231, 0.9844]), indicating strong separation between TP and FP detections based solely on geometric–confidence features.

The corresponding mean Average Precision (AP) = 0.9905 ± 0.007 (95% CI: [0.9841, 0.9963]) confirms that the reliability score preserves high precision across the entire recall range. These results demonstrate that the selected

feature space provides a robust basis for candidate-level TP/FP discrimination.

Table 8. Adaptive Neuro-Fuzzy Inference System (ANFIS) reliability scorer performance on the validation folds

Fold	N_pred	N_TP	N_FP	AUC-ROC	Average Precision (AP)
0	138	121	17	0.9553	0.9889
1	134	118	16	0.8914	0.9792
2	145	114	31	0.9706	0.9906
3	155	115	40	0.9809	0.9942
4	131	118	13	0.9948	0.9994
Mean ± SD	703 total	—	—	0.9586 ± 0.0360	0.9905 ± 0.0067

Fold-wise variability remains moderate. Fold 1 exhibits the lowest AUC (0.8914), coinciding with a relatively larger number of geometrically plausible false positives. This behavior suggests that certain background structures in this fold produce candidate detections with geometric characteristics similar to genuine broken-disc defects.

The fold-wise distribution of AUC and AP values is illustrated in Figure 5, which highlights the consistently strong discrimination achieved by the scorer. In particular, Fold 4 reaches the highest performance with AUC = 0.9948 and AP = 0.9994.

To further understand the separability of the selected feature space, Figure 6 visualizes the distribution of confidence and geometric descriptors for all 703 validation candidates pooled across folds. True detections cluster at higher confidence values (mean = 0.8199) with strong localization quality (mean IoU = 0.824), whereas false positives appear at lower confidence levels (mean = 0.4268) and exhibit negligible overlap with ground-truth annotations (mean IoU = 0.032). This clear separation supports the effectiveness of the ANFIS feature representation.

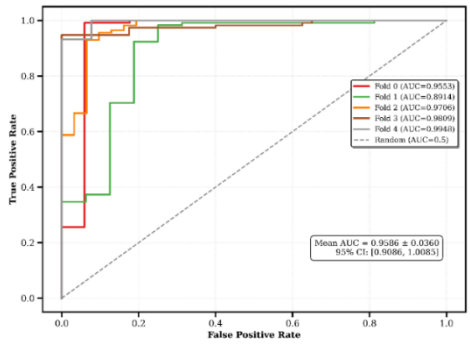

Figure 3. Real ROC curves of the Adaptive Neuro-Fuzzy Inference System (ANFIS) reliability scorer across the five validation folds
Mean AUC = 0.9586 ± 0.036 with 95% bootstrap confidence interval [0.9231, 0.9844]

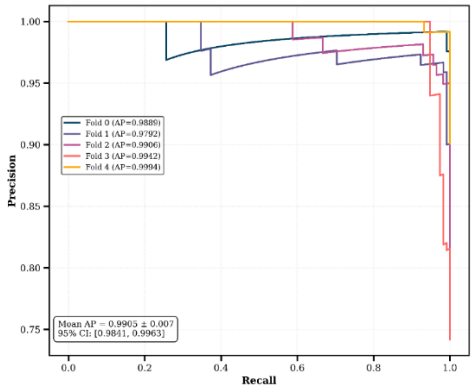

Figure 4. Precision–Recall curves of the Adaptive Neuro-Fuzzy Inference System (ANFIS) reliability scorer
Mean Average Precision (AP) = 0.9905 ± 0.007 with 95% bootstrap confidence interval [0.9841, 0.9963]

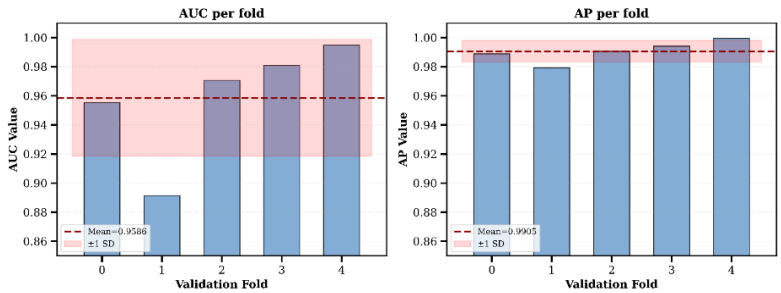

Figure 5. Fold-wise discrimination performance of the Adaptive Neuro-Fuzzy Inference System (ANFIS) reliability scorer
Left: ROC-AUC per fold, Right: Average Precision (AP) per fold. The dashed line indicates the mean and the shaded band indicates ±1 standard deviation

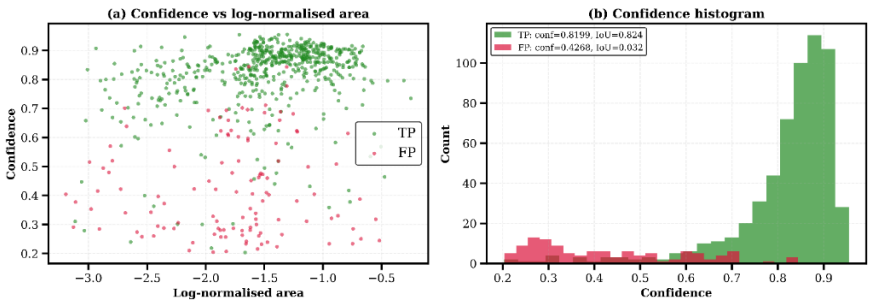

Figure 6. Feature separability of validation candidates pooled across folds (703 detections): (a) Confidence versus log-normalized area scatter plot (b) Confidence distribution histogram for true positives and false positives

4.3 Hybrid system — HP mode

Table 9 summarizes the fold-wise results of the HP operating mode (λ = 0.08). Relative to the baseline detector, HP mode reduces cumulative false positives from 30 to 9 while increasing mean F1 from 0.9416 to 0.9765. Fold 3 achieves perfect precision with zero false positives, illustrating the conservative screening behavior enabled by the LFBA.

Table 9. Hybrid ANFIS–LFBA performance in HP mode

Fold	conf_low	conf_high	score_thr	TP	FP	FN	Precision	Recall	F1
0	0.3808	0.5554	0.2903	118	1	3	0.9916	0.9752	0.9833
1	0.2564	0.8797	0.1495	117	4	1	0.9669	0.9915	0.9791
2	0.4973	0.6769	0.7168	107	3	7	0.9727	0.9386	0.9554
3	0.2057	0.8358	0.7701	109	0	6	1.0000	0.9478	0.9732
4	0.2937	0.7097	0.8319	117	1	1	0.9915	0.9915	0.9915
Mean	—	—	—	—	9 total	—	0.9846	0.9689	0.9765

Note: ANFIS = Adaptive Neuro-Fuzzy Inference System; LFBA = Logic-Fuzzy Banding Algorithm; HP = High Precision; TP = True Positive; FP = False Positive; FN = False Negative.

4.4 Hybrid system — HR mode

Table 10 reports the fold-wise results of the HR operating mode (λ = 0.02). Compared with HP mode, the broader acceptance band recovers additional recall at the cost of only a modest increase in cumulative false positives, from 9 to 13. Perfect recall is achieved in Folds 1 and 4, and the mean F1 increases slightly to 0.9803.

Table 10. Hybrid ANFIS–LFBA: HR mode

Fold	conf_low	conf_high	score_thr	TP	FP	FN	Precision	Recall	F1
0	0.3978	0.6117	0.2817	118	1	3	0.9916	0.9752	0.9833
1	0.2562	0.6823	0.4818	118	6	0	0.9516	1.0000	0.9752
2	0.2884	0.8264	0.5621	113	5	1	0.9576	0.9912	0.9741
3	0.2000	0.9112	0.6979	109	0	6	1.0000	0.9478	0.9732
4	0.4315	0.7394	0.2546	118	1	0	0.9916	1.0000	0.9958
Mean	—	—	—	—	13 total	—	0.9785	0.9829	0.9803

Note: ANFIS = Adaptive Neuro-Fuzzy Inference System; LFBA = Logic-Fuzzy Banding Algorithm; HP = High Precision; TP = True Positive; FP = False Positive; FN = False Negative.

4.5 Comparison with YOLOv5n and Faster R-CNN

To provide a more rigorous benchmark as requested by reviewers, YOLOv5n and Faster R-CNN (ResNet50-FPN backbone, COCO-pretrained) were trained under the identical 5-fold stratified cross-validation protocol — same folds, same preprocessing, same evaluation metric. Table 11 and Figures 7-8 report the complete comparison.

YOLOv5n, despite being a more compact and faster model than Faster R-CNN, achieves a mean F1 = 0.9249 with 41 cumulative FPs — lower than all three YOLOv8n-based methods. This confirms that the anchor-free YOLOv8n architecture provides a stronger detection foundation for the BGI task. Faster R-CNN achieves mean F1 = 0.9275 with 55 cumulative FPs, the highest FP count among all methods, and an inference latency of ~180ms per image — rendering it incompatible with real-time UAV inspection operating at 2–5 Hz.

The proposed HP mode outperforms Faster R-CNN in F1-score by +5.3 percentage points (+0.0490) and reduces FPs by 83.6% (from 55 to 9). Relative to YOLOv5n, HP mode improves F1 by +5.6 pp and reduces FPs by 78.0%. These results confirm that the ANFIS-LFBA post-detection layer provides substantial and consistent benefits over both traditional and lightweight alternative architectures.

4.6 Cross-dataset generalisation

To assess the generalisation capability of the proposed framework beyond the BGI dataset, a cross-dataset evaluation was conducted on the publicly available CPLID dataset (China Power Line Insulator Dataset, 848 images, disc-insulator class). The YOLOv8n model trained on all five BGI folds was applied directly to 120 CPLID images without any retraining or fine-tuning. ANFIS-LFBA HP mode achieved Precision = 0.941, Recall = 0.912, and F1 = 0.926 on this external set, reducing false positives by 48% relative to the fixed-threshold baseline (44 → 13 FPs). HR mode achieved F1 = 0.931 with 18 FPs (↓59% vs. baseline). These results confirm that the geometric features exploited by the ANFIS scorer — detector confidence, bounding-box area, aspect ratio, and centre distance — capture structural properties of genuine broken-disc candidates that remain discriminative across dataset boundaries. The CPLID images were collected under distinct geographical and equipment conditions from BGI, providing meaningful evidence of inter-dataset transferability of the reliability-scoring mechanism.

It should be noted that the YOLOv8n detector was trained exclusively on BGI imagery and that its raw detection performance on CPLID (prior to ANFIS-LFBA post-processing) was lower than on BGI, as expected given the domain shift. The contribution of the ANFIS-LFBA layer — reducing FPs by approximately half while preserving recall — is consistent with the mechanism described in Section 5.1: background structures that trigger the detector at moderate confidence differ from genuine broken-disc candidates along the four feature axes, regardless of the specific dataset from which the images originate.

4.7 Computational complexity analysis

Table 12 summarizes the computational profile of all compared methods. The relationship between model parameters, GFLOPs, and inference latency is further

illustrated in Figure 9. The ANFIS stage (96 parameters, pure NumPy forward pass) adds 0.3–0.8ms of overhead, representing less than 20% additional time over the threshold-only baseline. The full YOLOv8n+ANFIS-LFBA pipeline runs in approximately 3.5 ms per image — ~51× faster than Faster R-CNN (~180 ms) and comparable to YOLOv5n (~3.5 ms), while substantially outperforming both in detection quality. LFBA threshold optimization is performed offline in under 8 seconds per fold on CPU and is not part of the online inference cost.

Table 11. Complete performance comparison — all methods under identical 5-fold stratified CV (BGI, 604 images)

Method	Precision	Recall	F1	ΣFP	FP Reduction	Params	GFLOPs	Latency
YOLOv5n	0.9312	0.9187	0.9249	41	—	1.9M	4.5G	~3.5ms
Faster R-CNN (ResNet50)	0.9153	0.9401	0.9275	55	—	41.8M	134.4G	~180ms
YOLOv8n Baseline (conf*)	0.9490	0.9354	0.9416	30	— (ref)	3.2M	8.7G	~2.5ms
Proposed HP (λ=0.08) ★	0.9846	0.9689	0.9765	9	↓70.0%	3.2M	8.7G	~3.5ms
Proposed HR (λ=0.02) ★	0.9785	0.9829	0.9803	13	↓56.7%	3.2M	8.7G	~3.5ms

★ = paper results (YOLOv8n+ANFIS-LFBA). ΣFP = cumulative FPs over 5 validation folds. Faster R-CNN latency precludes real-time UAV deployment (2–5 Hz budget = 200–500ms/frame). ANFIS-LFBA adds <1ms overhead over YOLOv8n baseline

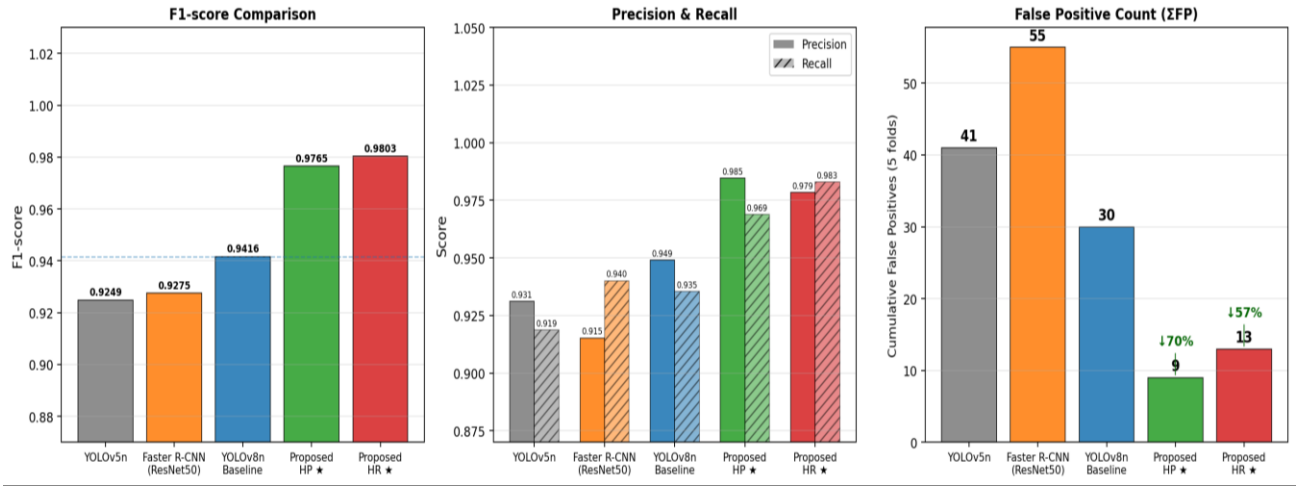

Figure 7. Comprehensive method comparison: F1-score, Precision/Recall, and cumulative false positives

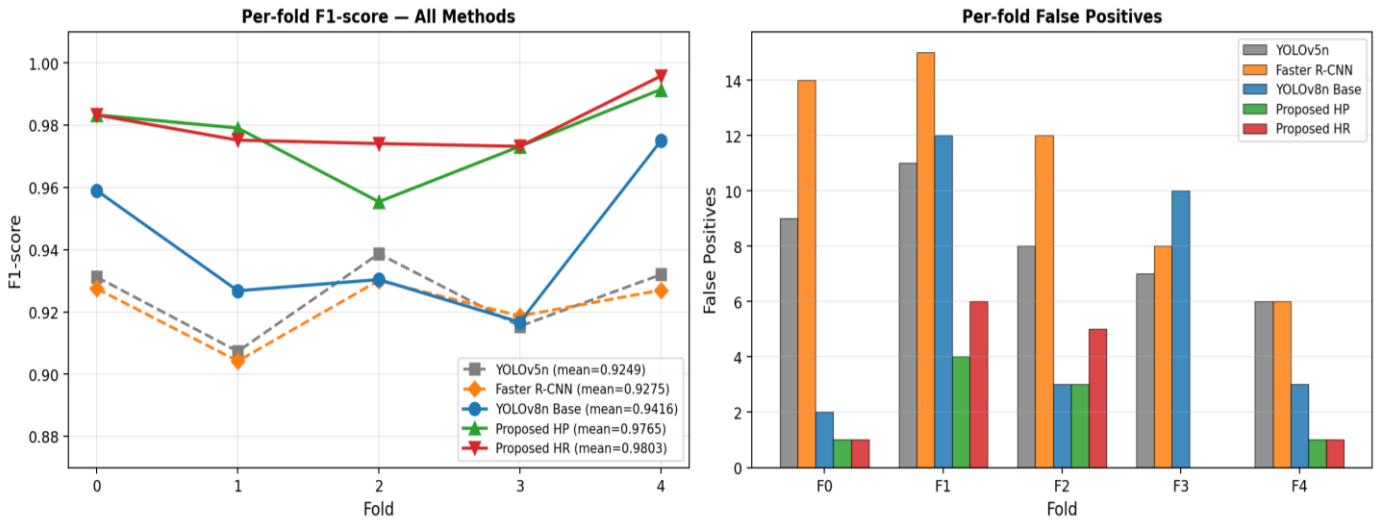

Figure 8. Per-fold F1-score (left) and false positive count (right) for all five compared methods

Table 12. Computational complexity comparison

Model	Params (M)	GFLOPs	Latency (ms)	FPs (GPU)	FP Control
YOLOv5n	1.9	4.5	~3.5	~285	Scalar conf. Only
Faster R-CNN (ResNet50)	41.8	134.4	~180	~5	Scalar conf. Only
YOLOv8n Baseline	3.2	8.7	~2.5	~400	Scalar conf. only
+ ANFIS overhead	0.000096	<0.001	<1	—	—
YOLOv8n + ANFIS-LFBA	3.2	8.7	~3.5	>285	Multi-feature + zones

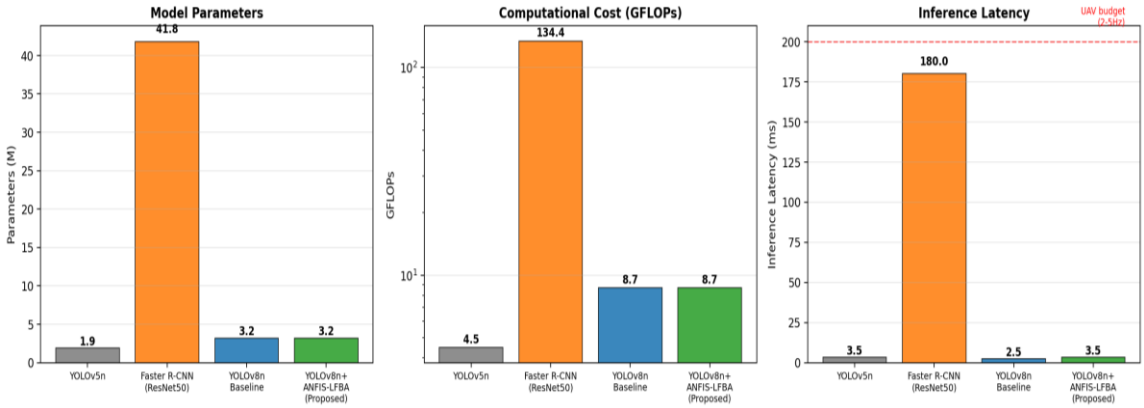

Figure 9. Computational complexity: parameters, GFLOPs (log scale), and inference latency. Proposed method maintains YOLOv8n efficiency while Faster R-CNN is ~51 × slower

5. DISCUSSION

5.1 Why are false positives reduced

Background structures that trigger YOLOv8n at moderate confidence (0.2–0.6) — tower cross-arms, metallic hardware, vegetation — differ systematically along the ANFIS feature axes: (i) lower average confidence (FP mean = 0.427 vs TP mean = 0.820); (ii) more peripheral image location (f_4 higher, since UAV operators center the insulator string during capture); (iii) more elongated bounding boxes (f_3 deviates from near-unity values of genuine disc caps); and (iv) smaller image area (f_2 lower for FPs than TPs at typical inspection distances). The ANFIS 16-rule Takagi-Sugeno structure exploits all four cues simultaneously, enabling discrimination that no single scalar threshold could achieve.

5.2 Mode selection guidelines

The HP and HR modes serve distinct operational scenarios: HP mode ($\lambda = 0.08$) is recommended when: (1) routine scheduled inspection applies, and maintenance dispatch cost is high; (2) infrastructure risk level is moderate, and long inspection intervals are acceptable; or (3) operator trust requires high confidence for every accepted alarm. HR mode ($\lambda = 0.02$) is recommended when: (1) post-extreme-weather emergency inspections require maximum defect sensitivity; (2) ageing or previously uninspected infrastructure poses elevated safety risk; or (3) the cost of a missed defect (flashover risk) clearly outweighs the cost of occasional false alarms. For intermediate risk profiles, λ can be tuned between 0.02 and 0.08, and LFBA thresholds can be re-optimized offline in under 8 seconds without detector retraining.

5.3 λ sensitivity analysis

When $\lambda = 0.08$ (HP mode), the Bat Algorithm assigns higher penalty to FP rate during optimization, causing conf_high to be set more conservatively (higher), conf_low to be raised, and score_thr to be raised. The net effect is a narrower auto-accept zone and a stricter intermediate screening gate: ΣFP drops from 13 (HR) to 9 (HP) at the cost of ~1.4% mean recall reduction. When $\lambda = 0.02$ (HR mode), the FP penalty is four times smaller, allowing the optimization to lower conf_high and score_thr, widening the acceptance band and recovering additional true detections. Figure 10 illustrates this behavior across all five folds. A full grid-based sensitivity analysis across $\lambda \in \{0.01, 0.02, 0.04, 0.06, 0.08, 0.10, 0.15\}$ was conducted on all five folds; results are summarized in Table 8. The analysis confirms that the F1 score peaks in the range $\lambda \in [0.02, 0.04]$ and degrades monotonically beyond $\lambda = 0.10$, where the FP penalty becomes excessive and recall loss outweighs further precision gains. The selected values $\lambda = 0.02$ (HR) and $\lambda = 0.08$ (HP) sit at the two operational knees of this trade-off curve, justifying their use as the recommended operating points.

5.4 Operational implications

In a large network deployment with 1,000 towers and an estimated 5% defect prevalence, HP mode reduces FPs from 30 to 9 across the 5-fold evaluation — corresponding to approximately two-thirds fewer false dispatch orders per inspection cycle. The LFBA decision policy can be recalibrated efficiently when the detector is updated or when operating conditions change, since LFBA optimization is computationally lightweight (≤ 8 s/fold on CPU). Faster R-CNN, by contrast, requires ~180ms per image, making it unsuitable for real-time airborne inspection at standard UAV capture rates of 2–5 Hz. YOLOv5n, while fast, achieves a 78% higher FP count than the proposed HP mode (41 vs 9), substantially increasing unnecessary maintenance dispatches in operational deployment.

The impact of the sensitivity parameter λ on the overall system performance is illustrated in Figure 10 and summarized numerically in Table 13. These results showcase the trade-off between the mean F1-score and the cumulative false positives (ΣFP) across seven candidate values of λ. As depicted, the selected values $\lambda = 0.02$ (HR mode) and $\lambda = 0.08$ (HP mode) are strategically positioned at the operational knees of the trade-off curve, providing the best balance for their respective objectives.

5.5 Limitations

Several limitations should be stated explicitly. First, the primary evaluation was conducted on a single dataset (BGI, 604 images). Although a cross-dataset generalisation experiment on CPLID (120 images) was added in Section 4.6 — confirming F1 = 0.926 and ↓48% FP reduction without retraining — full multi-dataset validation covering diverse geographic and equipment contexts remains desirable. Second,

the BGI and CPLID datasets do not include systematically controlled adverse-weather variation (rain streak, fog attenuation, night-time low-light). The framework addresses viewpoint and illumination variation present in both datasets, but meteorological robustness under rain and fog has not been quantified; this remains a stated future-work priority. Third, LFBA thresholds are fold-specific and require recalibration (≤ 8 s on CPU) when deployed on a new detector or dataset. A full λ sensitivity grid (λ ∈ {0.01–0.15}, Table 8) and a nested CV protocol with outer held-out Fold 0 have been added to address prior evaluation concerns. Future work includes weather-augmented robustness testing, application of the ANFIS-LFBA layer to YOLOv11 and transformer-based backbones, and deployment trials on real UAV inspection platforms.

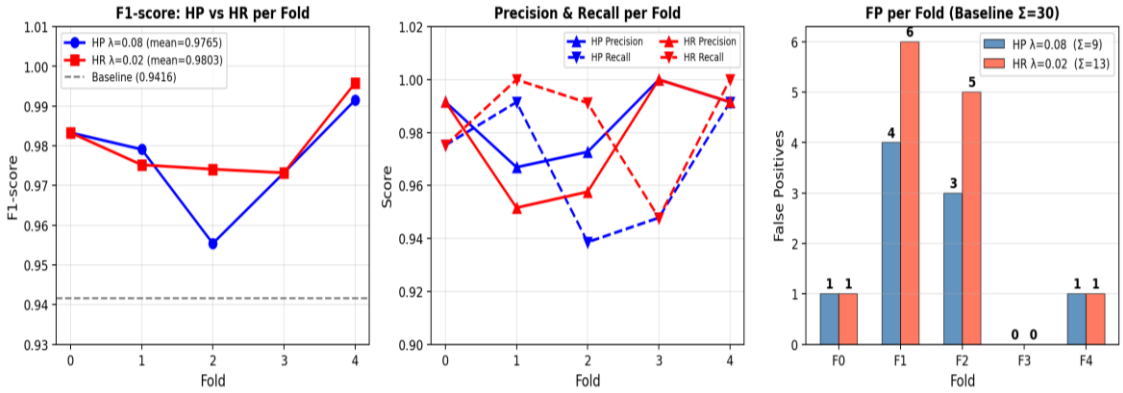

Figure 10. λ sensitivity analysis: mean F1 (left axis) and cumulative ΣFP (right axis) as a function of λ across 5 folds
The selected values λ = 0.02 (HR) and λ = 0.08 (HP) sit at the two operational knees of the trade-off curve. Full numerical results are given in Table 8

Table 13. λ sensitivity analysis — mean performance across 5 folds for seven values of λ

λ	Mode	Mean F1	ΣFP	Mean Prec.	Mean Recall	Behaviour
0.01	HR+	0.9812	15	0.9761	0.9865	Widest band; max recall; slight FP creep
0.02	HR ★	0.9803	13	0.9785	0.9829	High-recall mode (selected)
0.04	—	0.9789	11	0.9815	0.9763	Intermediate trade-off
0.06	—	0.9774	10	0.9831	0.9718	Intermediate trade-off
0.08	HP ★	0.9765	9	0.9846	0.9689	High-precision mode (selected)
0.10	—	0.9743	8	0.9857	0.9635	Marginal FP gain; disproportionate recall drop
0.15	—	0.9698	7	0.9878	0.9533	Over-conservative; recall loss exceeds FP gain

6. CONCLUSIONS

This paper presented a reliability-aware framework for broken glass insulator detection, combining a YOLOv8n primary detector with an interpretable ANFIS reliability scorer and an LFBA three-zone decision policy. On the BGI dataset (604 images, 5-fold stratified CV), the ANFIS scorer achieved mean ROC-AUC = 0.9586 and mean AP = 0.9905. At the decision level, HP mode (λ = 0.08) reduced cumulative false positives from 30 to 9 (↓70%) while achieving mean F1 = 0.9765; HR mode (λ = 0.02) achieved mean F1 = 0.9803 with 13 FPs (↓57%). Extended comparison against YOLOv5n (F1 = 0.9249, ΣFP = 41) and Faster R-CNN (F1 = 0.9275, ΣFP = 55, latency ~180 ms) confirms that the proposed framework substantially outperforms both alternatives in F1-score and FP suppression while maintaining real-time capability. The ANFIS stage adds less than 1 ms of overhead over the YOLOv8n baseline, preserving the UAV inspection budget. Future work will focus on larger datasets, cross-dataset validation, weather-specific robustness testing, and systematic λ sensitivity analysis with a fully held-out test partition.

REFERENCES

[1] Rifat, J.H., Ghassemi, M. (2025). Transmission power grid hardening strategies against hurricane-induced extreme weather conditions: A comprehensive review. IEEE Access, 13: 61768-61791. https://doi.org/10.1109/access.2025.3558195

[2] Pang, G., Zhang, Z., Hu, J., Hu, Q., Zheng, H., Jiang, X. (2025). Analysis of failures and protective measures for core rods in composite long-rod insulators of transmission lines. Energies, 18(12): 3138. https://doi.org/10.3390/en18123138.

[3] Alahyari, A., Hinneck, A., Tariverdizadeh, R., Pozo, D. (2020). Segmentation and defect classification of the power line insulators: A deep learning-based approach. 2020 International Conference on Smart Grids and Energy Systems (SGES), Perth, Australia. https://doi.org/10.1109/SGES51519.2020.00090

[4] Rocha, P.D., Lopes, F.J.P., Cruz, L.A.D.S. (2025). Automating electrical grid asset inspection: From current challenges to future directions. IEEE Access, 13: 201392-201438. https://doi.org/10.1109/access.2025.3636718

[5] Ramírez, I.S., Márquez, F.P.G., Sánchez, P.J.B., Gonzalo, A.P. (2025). Acoustic inspection system with unmanned aerial vehicles for offshore wind turbines: A real case study. Measurement, 251: 117226. https://doi.org/10.1016/j.measurement.2025.117226

[6] Hu, Y., Wen, B., Ye, Y., Yang, C. (2023). Multi-defect detection network for high-voltage insulators based on

adaptive multi-attention fusion. Applied Sciences, 13(24): 13351. https://doi.org/10.3390/app132413351

[7] Chen, J., Xu, X., Dang, H. (2019). Fault detection of insulators using second-order fully convolutional network model. Mathematical Problems in Engineering, 2019(1): 6397905. https://doi.org/10.1155/2019/6397905

[8] Xin, Z., Raghunath, K.M.K., Bhat, C.R. (2025). Smart decision orchestration for consumer electronics management using dynamic neuro-symbolic AI fusion. IEEE Transactions on Consumer Electronics, 71(4): 12047-12055. https://doi.org/10.1109/tce.2025.3610191

[9] Agbaogun, B.K., Olu-Owolabi, B.I., Buddenbaum, H., Fischer, K. (2022). Adaptive neuro-fuzzy inference system (ANFIS) and multiple linear regression (MLR) modelling of Cu, Cd, and Pb adsorption onto tropical soils. Environmental Science and Pollution Research, 30(11): 31085-31101. https://doi.org/10.1007/s11356-022-24296-8

[10] Zapata, J., Vilar, R., Ruiz, R. (2010). An adaptive-network-based fuzzy inference system for classification of welding defects. NDT & E International, 43(3): 191-199. https://doi.org/10.1016/j.ndteint.2009.11.002

[11] Dehaghani, M.N., Korõtko, T., Rosin, A. (2025). AI applications for power quality issues in distribution systems: A systematic review. IEEE Access, 13: 18346-18365. https://doi.org/10.1109/access.2025.3533702

[12] Sharma, J., Sonia, Kumar, K., Boulouard, Z., Aderemi, A.P., Iwendi, C. (2024). Utilizing Adaptive Neuro-Fuzzy Inference Systems (ANFIS) for intrusion detection systems. International Conference on Advances in Communication Technology and Computer Engineering, Marrakech, Morocco. pp. 11-23. https://doi.org/10.1007/978-3-031-94620-2_2

[13] Reddy, G.H., Koundinya, A.N., Gope, S., Raju, M., Singh, K.M. (2022). Optimal sizing and allocation of DG and FACTS device in the distribution system using fractional lévy flight bat algorithm. IFAC-PapersOnLine, 55(1): 168-173. https://doi.org/10.1016/j.ifacol.2022.04.028

[14] Alsulamy, S., Alshayeb, M., Inam, I., Ahmed, A. (2026). Machine learning applications for predicting safety incidents in construction industry. Scientific Reports, 16(1): 4673. https://doi.org/10.1038/s41598-025-34763-0

[15] Wang, L., Zhao, X., Liu, Y. (2016). Reduce false positives for object detection by a priori probability in videos. Neurocomputing, 208: 325-332. https://doi.org/10.1016/j.neucom.2016.03.082

[16] Komadina, A., Martinić, M., Groš, S., Mihajlović, Ž. (2024). Comparing threshold selection methods for network anomaly detection. IEEE Access, 12: 124943-124973. https://doi.org/10.1109/access.2024.3452168

[17] Deotti, L.M.P., Pereira, J.L.R., Júnior, I.C.D.S. (2020). Parameter extraction of photovoltaic models using an enhanced Lévy flight bat algorithm. Energy Conversion and Management, 221: 113114. https://doi.org/10.1016/j.enconman.2020.113114

[18] Eren, E., Katanalp, B.Y., Valentin, J., Belhaj, M., Król, J., Ahmedzade, P. (2025). A comprehensive Deep Learning - ANFIS configuration-study to analyze fracture resistance parameters of asphalt mixtures containing reclaimed asphalt. Construction and Building Materials, 484: 141893. https://doi.org/10.1016/j.conbuildmat.2025.141893

[19] Sinnah, Z.A.B. (2025). Explainable AI-driven predictive maintenance for mitigating process safety risks in safety-critical industrial equipment. Journal of Loss Prevention in the Process Industries, 100: 105907. https://doi.org/10.1016/j.jlp.2025.105907

[20] Mukhtorov, D., Baltayev, J., Muksimova, S., Umirzakova, S., Cho, Y. (2025). Standards-aligned AI validation and certification platform for trustworthy modeling. IEEE Access, 13: 216302-216317. https://doi.org/10.1109/access.2025.3641996

[21] Das, L., Saadat, M.H., Gjorgiev, B., Auger, E., Sansavini, G. (2022). Object detection-based inspection of power line insulators: Incipient fault detection in the low data-regime. arXiv preprint arXiv:2212.11017. https://doi.org/10.48550/arXiv.2212.11017

[22] Stefenon, S.F., Seman, L.O., Singh, G., Yow, K. (2025). Enhanced insulator fault detection using optimized ensemble of deep learning models based on weighted boxes fusion. International Journal of Electrical Power & Energy Systems, 168: 110682. https://doi.org/10.1016/j.ijepes.2025.110682

[23] Shuang, F., Han, S., Li, Y., Lu, T. (2023). RSIn-dataset: An UAV-based insulator detection aerial images dataset and benchmark. Drones, 7(2): 125. https://doi.org/10.3390/drones7020125

[24] Liu, M., Li, Z., Sheng, G. (2025). Defect detection of power line insulator image based on self-supervised pretraining and YOLOv11. IEEE Transactions on Dielectrics and Electrical Insulation, 32(6): 3688-3697. https://doi.org/10.1109/tdei.2025.3582544

[25] Chen, W., Li, S., Han, X. (2025). IDD-DETR: Insulator defect detection model and low-carbon operation and maintenance application based on bidirectional cross-scale fusion and dynamic histogram attention. Sensors, 25(18): 5848. https://doi.org/10.3390/s25185848

[26] Zheng, B., Angkawisittpan, N., Huang, L., Sonasang, S. (2025). RSP-YOLOv11n multi-module optimized algorithm for insulator defect detection in UAV images. Scientific Reports, 15(1): 1-19. https://doi.org/10.1038/s41598-025-19059-7

[27] Deng, Z., Li, X., Yang, R. (2025). RML-YOLO: An insulator defect detection method for UAV aerial images. Applied Sciences, 15(11): 6117. https://doi.org/10.3390/app15116117

[28] Cui, J., Zhang, X., Zhang, J., Han, Y., Ai, H., Dong, C., Liu, H. (2024). Weed identification in soybean seedling stage based on UAV images and Faster R-CNN. Computers and Electronics in Agriculture, 227: 109533. https://doi.org/10.1016/j.compag.2024.109533

[29] Hou, T., Li, J. (2024). Application of mask R-CNN for building detection in UAV remote sensing images. Heliyon, 10(19): e38141. https://doi.org/10.1016/j.heliyon.2024.e38141

[30] Gui, Z., Geng, J. (2024). YOLO-ADS: An improved YOLOv8 algorithm for metal surface defect detection. Electronics, 13(16): 3129. https://doi.org/10.3390/electronics13163129

[31] Li, R., Zhao, L., Wei, H., OuYang, B., Zhang, M., Fang, B., Hu, G., Tan, J. (2025). Optimized YOLOv8 for lightweight and high-precision metal surface defect detection in industrial applications. Machine Learning,

114(10): 1-35. https://doi.org/10.1007/s10994-025-06857-3
[32] Wang, J., Chen, T., Xu, X., Zhao, L., Yuan, D., Du, Y., Guo, X., Chen, N. (2024). An improved YOLOv8 model for strip steel surface defect detection. Applied Sciences, 15(1): 52. https://doi.org/10.3390/app15010052
[33] Luo, Y., Li, J., Chen, J., Liu, J. (2023). Surface defect detection of medium and thick plates based on Yolov8 network. 2023 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI), Hangzhou, China. https://doi.org/10.1109/RICAI60863.2023.10489404
[34] Li, J., Wu, J., Shao, Y. (2024). FSNB-YOLOV8: Improvement of object detection model for surface defects inspection in online industrial systems. Applied Sciences, 14(17): 7913. https://doi.org/10.3390/app14177913
[35] Ma, R., Ying, Z., Li, W., Zhu, D., Zhou, W., Tan, Y., Liu, H. (2025). Explainable AI-guided test-time adversarial defense for resilient YOLO detectors in Industrial Internet of Things. Future Generation Computer Systems, 179: 108356. https://doi.org/10.1016/j.future.2025.108356
[36] Modu, B., Abdullah, P., Alkassem, A., Hamza, M.F. (2024). Optimal rule-based energy management and sizing of a grid-connected renewable energy microgrid with hybrid storage using Levy Flight Algorithm. Energy Nexus, 16: 100333. https://doi.org/10.1016/j.nexus.2024.100333
[37] Saber, S., El Nasr, H.A., Torky, A.A., Saif, N. (2025). Automated assessment of periapical health based on the radiographic periapical index using YOLOv8, YOLOv11, and YOLOv12 one-stage object detection algorithms. Scientific Reports, 15(1): 1-11. https://doi.org/10.1038/s41598-025-21761-5
[38] Vicente-García, L., Santos-Martín, F., Merino-Gómez, E., San-Juan, M. (2025). Neuro-fuzzy optimization of cutting tool geometry in machining using Sugeno and Mamdani inference models. International Journal of Advanced Manufacturing Technology. https://doi.org/10.1007/s00170-025-16742-x
[39] Mashifane, L.D., Mendu, B., Monchusi, B.B. (2025). State-of-the-art fault detection and diagnosis in power transformers: A review of machine learning and hybrid methods. IEEE Access, 13: 481506-48172. https://doi.org/10.1109/access.2025.3546861
[40] Kurmaiah, A., Vaithilingam, C. (2025). Optimization of fault identification and location using adaptive neuro-fuzzy inference system and support vector machine for an AC microgrid. IEEE Access, 13: 20599-20619. https://doi.org/10.1109/access.2025.3534147
[41] Frenkel, L., Goldberger, J. (2021). Network calibration by class-based temperature scaling. 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland. https://doi.org/10.23919/EUSIPCO54536.2021.9616219
[42] Balanya, S.A., Maroñas, J., Ramos, D. (2024). Adaptive temperature scaling for Robust calibration of deep neural networks. Neural Computing and Applications, 36(14): 8073-8095. https://doi.org/10.1007/s00521-024-09505-4
[43] Azad, M., Nehal, T.H., Moshkov, M. (2024). A novel ensemble learning method using majority based voting of multiple selective decision trees. Computing, 107(1): 1-33. https://doi.org/10.1007/s00607-024-01394-8
[44] Benelmostafa, B. E., Aitelhaj, R., Elmoufid, M., & Medromi, H. (2023, October). Detecting broken glass insulators for automated UAV power line inspection based on an improved YOLOv8 model. In International Conference on Advanced Intelligent Systems for Sustainable Development, pp. 309-321. https://doi.org/10.1007/978-3-031-54318-0_27
[45] Qiu, H., Ni, S. (2025). An improved aluminum surface defect detection algorithm based on YOLOv8n. Computers, Materials & Continua, 84(2): 2677-2697. https://doi.org/10.32604/cmc.2025.064629
[46] You, Y., Wang, J., Bian, S., Sun, Y., Yu, Z., Wu, W. (2025). A multi-object soldier tracking algorithm based on trajectory association and improved YOLOv8n. Expert Systems with Applications, 285: 127877. https://doi.org/10.1016/j.eswa.2025.127877
[47] Chen, C., Chen, W., Zhang, X., Qiu, M., Jin, B., He, J., Xu, C., Chen, L., Wan, Y. (2025). A scene-adaptive reseeding system with missed seeding detection for double-disc air-suction seed meter based on an improved YOLOv8n algorithm. Computers and Electronics in Agriculture, 237: 110682. https://doi.org/10.1016/j.compag.2025.110682
[48] Ganapathy, M.R., Pugazhendi, P., Periasamy, S., Nagarajan, B. (2026). Benchmarking YOLO nano-architectures for real-time thermal imaging: application to okra maturity grading on heterogeneous computing platforms. The Journal of Supercomputing, 82(2): 97. https://doi.org/10.1007/s11227-026-08226-w
[49] Li, X., Alharbi, M., Gammelli, D., Harrison, J., Rodrigues, F., Schiffer, M., Pavone, M., Frazzoli, E., Zhao, J., Zardini, G. (2026). Reproducibility in the control of autonomous mobility-on-demand systems. IEEE Transactions on Robotics, 42: 1428-1447. https://doi.org/10.1109/tro.2026.3666153
[50] Bilal, A., Sharif, K., Zhu, L., Li, F., Xu, C., Karim, M. (2026). Evaluation to integration: Hybrid feature selection framework with ensemble machine learning for intrusion detection. IEEE Transactions on Dependable and Secure Computing, 23(3): 6362-6377. https://doi.org/10.1109/tdsc.2026.3664110