Results in Engineering 29 (2026) 108833

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

From vineyard to vision: Multi-domain analysis and mitigation of grape cluster detection failures in complex viticultural environments

Shubham Rana [a,*], Oliver Hensel [a], Abozar Nasirahmadi [a,b]

[a] Department of Agricultural and Biosystems Engineering, University of Kassel, Witzenhausen, D-37213, Germany
[b] Department of Energy and Technology, Swedish University of Agricultural Sciences, Box 7032, Uppsala 75007, Sweden

ARTICLE INFO

Keywords:
Cross domain model generalization
Perception reliability
Autonomous vineyard robotics
Spectral domain-aware training
Multispectral near-infrared data analytics

ABSTRACT

Accurate and robust grape-cluster detection remains a persistent challenge in precision viticulture due to spectral variability, canopy occlusion, and lighting heterogeneity. Recent advancements in the YOLO series, have focused on eliminating post-processing bottlenecks like Non-Maximum Suppression (NMS) to improve inference speed. Furthermore, state-of-the-art models increasingly integrate attention-based mechanisms and hybrid transformer-CNN backbones to enhance feature representation and global context understanding, leading to greater accuracy. This study presents a comprehensive benchmark and error analysis of recent YOLO architectures (v8–v12), including an orientation-aware YOLOv8-OBB, where YOLOv11 and YOLOv12 are community implementations rather than official successors to the Ultralytics, across multispectral (RGB, NIR) vineyard datasets under both normal and degraded imaging conditions. Models were evaluated using standard metrics (Precision, Recall, F1, mAP@0.5, mAP@0.5:0.95) and False Classification Rate (FCR) that integrates false positives and negatives to capture field reliability. Results show that YOLOv10, YOLOv11, and YOLOv8-OBB deliver the highest overall stability and transfer performance, maintaining consistent F1 $\geq$ 0.85 across spectral regimes. RGB imagery outperforms NIR by approximately 8–10%, yet OBB regression markedly improves NIR localization, reducing FCR by up to 30% in poor-quality scenes. Cross-dataset experiments further reveal that YOLOv11 sustains the lowest metric variance, while YOLOv8-OBB achieves superior mAP@0.5:0.95 when object orientations vary. The findings emphasize that orientation-aware geometry, domain-robust feature balance, and variance-based reliability metrics are more predictive of field performance than absolute mAP values. The study provides actionable guidance for detector selection in vineyard monitoring and establishes a reproducible benchmark for multispectral object detection under real-world variability.

1. Introduction

In modern viticulture, computer vision has become a critical technology for grape monitoring, enabling objective and high-throughput data collection that was previously infeasible [1]. The primary application is yield estimation, where deep learning models, such as YOLO, are trained to detect and count grape clusters from images captured by ground vehicles or drones, offering a precise forecast for harvest logistics [2]. Beyond counting, these vision systems are essential for quality and ripeness assessment. By analysing RGB images, algorithms can track visual phenotypes like berry colour, size, and cluster compactness [3]. More advanced hyperspectral and multispectral imaging systems can non-destructively estimate internal quality metrics, correlating spectral signatures with chemical compounds like soluble solids (Brix) and phenolic content [4]. Furthermore, computer vision is the engine of precision viticulture; UAVs equipped with multispectral and thermal cameras generate detailed maps of vine vigour (e.g., NDVI) and water stress, allowing growers to identify and manage in-field variability [5].

This vision-based monitoring data serves as the perceptual foundation for automation and robotics in the vineyard. The most prominent application is robotic harvesting, where vision systems are not only required to detect and locate grape clusters in a complex, occluded 3D environment but also to identify the precise cutting point of the stem, also known as peduncle [1]. For automated maintenance, robots rely on 3D vision, using stereo or RGB-D cameras to reconstruct the woody cane structure of dormant vines, enabling intelligent and precise robotic pruning [6]. Similarly, vision-guided sprayers can detect the presence and density of foliage, allowing for targeted application of treatments

* Corresponding author.
E-mail address: shubham.rana@uni-kassel.de (S. Rana).

https://doi.org/10.1016/j.rineng.2025.108833
Received 9 November 2025; Received in revised form 11 December 2025; Accepted 20 December 2025
Available online 20 December 2025

only where needed, reducing chemical use [7]. This automation extends into the winery with automated optical sorting with grapes on a conveyor are scanned by high-speed cameras, and deep learning models classify each berry in real-time, allowing robotic air-jet actuators to remove unripe berries, raisins, or material other than grapes with high precision [8].

Recent advances confirm that computer vision and deep learning have transformed digital viticulture, providing reliable tools for automated monitoring, yield estimation, and targeted intervention. Deep learning-based frameworks such as YOLO, U-Net, and Mask R-CNN have been applied to grapevine images for yield estimation, disease identification, and precision operations, with growing adoption in commercial vineyards due to their scalability and non-destructive capabilities [9–13].

Despite these advances, reliable detection of grape bunches in real field conditions is still hampered by several factors such as

I) **Occlusion from leaves, trellis structures, and other bunches** has been consistently shown to be the primary source of false negatives and inaccuracy, requiring novel multi-stage or attention-based model designs to partially mitigate this effect [11–15].
II) **Drastic illumination and background variability** such as sun/shade transitions between row interiors and exteriors which further complicate detection, as most models experience a significant drop in detection accuracy under non-uniform lighting or when imaging grape clusters at varying times of day [15,16].
III) **Grape cluster morphology** (compact vs. loose, cultivar-specific variability) has also been highlighted as a major challenge, demanding robust, highly adaptable computer vision algorithms wherein imaging angle can additionally affect berry count accuracy, especially for complex clusters [12].
IV) **Generalization to diverse field environments** and across all phenological stages remains an unsolved issue, with current research emphasizing the need for larger, more varied annotated datasets and more context-aware models [14,17]

Object detection, especially real-time models, is pivotal to precision and digital agriculture, enabling in-field decisions for smart farming, automation, and monitoring tasks [18,19]. YOLO models are widely used in agricultural object detection owing to their competitive accuracy and real-time performance. These attributes make YOLO preferable over slower two-stage detectors. YOLO is therefore extensively used in agricultural equipment, UAV systems, and edge analytics for crop monitoring due to its balance of speed and accuracy [20,21]. Each new YOLO version provides architectural changes that improve speed, computational efficiency, or detection performance. YOLOv4 introduced big leaps in performance and deployment on commodity GPUs [22]. YOLOv7, YOLOv9, and YOLOv10 have continued this trend, with reports of increased mAP, reduced parameters, and new backbone/neck designs suitable for mobile and edge device [20–35]. These advances in YOLO technology support the application of real-time embedded devices essential for vineyard analytics.

Most current fruit- and bunch-detection systems in orchards and vineyards still rely on Horizontal Bounding Boxes (HBB), which degrade in performance when targets are rotated, elongated along trellis, densely packed, or partially occluded, because axis-aligned boxes capture excessive background and poorly approximate true cluster extent, leading to both duplicate and missed detections [23–27]. In contrast, Oriented Bounding Boxes (OBB) have repeatedly improved localization and reduced redundant detections for arbitrarily oriented, high–aspect-ratio objects in remote sensing and aerial imagery by aligning boxes to object pose and tightening support under occlusion [23,29–31,36]. Similar gains have been reported for non-axis-aligned structures in horticulture, including dense wheat spikes, where OBBs better capture instance extent and suppress background interference [30,32]. These properties are directly relevant to vine canopies, where grape clusters frequently appear rotated, elongated, and partially hidden behind leaves, so orientation-aware detection is expected to better handle aspect-ratio sensitivity and provide tighter localization than HBB in complex scenes. However, systematic OBB evaluations in viticulture remain scarce, particularly across regions that differ in canopy management, trellising systems, cultivar morphology, and imaging practices, even though such diversity is critical for assessing whether orientation-aware models generalize beyond a single vineyard or country. This work therefore proposes a multi-dataset benchmark that jointly compares HBB and OBB YOLO variants using standard accuracy metrics and an error-centric False Classification Rate, tests the hypothesis that orientation-aware architectures offer more reliable bunch localization in complex canopies, and quantifies how these gains translate into robustness for cross-region, cross-spectral vineyard analytics [26–33].

High-resolution RGB imagery often matches or exceeds the performance of multispectral images as its cost-effective deployment has made it standard in vineyard applications for monitoring, segmentation, and yield estimation [37–39]. Multispectral imaging and NIR bands provide increased contrast between plants and background, improving classification in challenging illumination, shadow, or adverse environmental conditions [37–42]. Multispectral vineyard segmentation studies show that NIR bands reliably enhance feature extraction in vine segmentation, especially under non-uniform illumination [38]. Studies demonstrated the use of NIR and multispectral imaging to improve grapevine detection under variable lighting and nighttime scenarios, with UAVs and other proximal sensing platforms [40]. Results indicate that while cross-spectral (RGB + NIR/multispectral) data sometimes improve detection or segmentation performance, it was task and crop dependent. In some cases, high-resolution RGB alone is adequate, and NIR adds little, while in others, NIR is critical [38,43]. Comparative studies between RGB vegetation indices and multispectral indices (NDVI, SAVI) show comparative performance for general monitoring, but more precise NIR-based indices may be needed for specific stress or disease assessments in vineyard scenarios [44]. Despite growth in sensor adoption, there is a significant lack of systematic, standardized comparative evaluations of RGB versus NIR/multispectral performance for grape bunch and disease detection in real-world vineyard field conditions; published experiments are often limited in sample size or generalizability. A systematic review in precision viticulture notes the scarcity of direct, controlled RGB versus NIR comparisons; the gap restricts validated protocol recommendations for practitioners [37].

Evaluation techniques also play a significant role in detecting imbalanced datasets typical in agriculture, particularly with clustered grapes where the costs of false positives and false negatives can diverge significantly especially in tasks like yield estimation and maturity assessment [45–47]. There is a problem of data imbalance with certain classes in viticultural datasets, like a specific grape type which can be underrepresented [14,48]. Traditional metrics such as mean Average Precision (mAP) may not sufficiently capture the intricacies of performance under such conditions [16,49]. So, we introduce the FCR as a compact reliability indicator particularly as the proportion of misclassified instances computed at a fixed IoU and confidence threshold, which complements precision, recall, and F1 by summarizing both false positives and false negatives in a single error-centric metric.

Robustness to real-world degradation factors such as blur, low illumination, and weather-induced noise remains one of the most under investigated challenges in vineyard computer vision, limiting the reliability of deployed models in practical settings [50]. Standard benchmarks like ImageNet-C and recent studies highlight how blur, noise, and environmental variability can drastically degrade neural network performance, yet adaptation of these corruption-aware frameworks to agricultural domains is still. Corruption-aware training strategies are designed for improved generalization under diverse, realistic corruptions and have demonstrated potential in broader computer vision but

require systematic translation to vineyard analytics for effective in-field deployment [50,51]. Grape-cluster detection in vineyards is also challenged by spectral overlap between clusters, foliage and soil, as well as strong intra-class heterogeneity in canopy structure and illumination. Similar overlap and heterogeneity effects have been reported more broadly in agricultural imagery, where soft-classification is needed to disentangle mixed pixels and spectrally similar vegetation–soil mixtures [41].

Meanwhile, public vineyard datasets such as WGISD [52] and CERTH [53] have become foundational for grape bunch detection research, offering carefully annotated imagery across multiple varieties and environments [52,53]. However, these datasets continue to display significant limitations. Most notably, their scale often remains insufficient for robust generalization, and their spectral range predominantly consists of RGB imagery, with few if any datasets systematically integrating multi- or hyperspectral bands needed for cross-spectral model comparisons [52,53]. Furthermore, existing vineyard datasets predominantly rely on HBB annotations - which, while standard, can fail to capture the geometry of rotated or elongated bunches whereas OBB annotations remain rare despite their clear potential to improve detection and localization accuracy in realistic, highly variable vineyard layouts [54].

The restricted diversity of currently available datasets is manifested both in terms of environmental variability and annotation comprehensiveness which altogether impede the development of universally applicable and robust computer vision models for viticulture [52,53]. This highlights an acute need for both larger, more cross-spectrally varied vineyard datasets and explicit inclusion of OBB annotations. Such expansions will support nuanced investigation of how modality choices like spectral bands and annotation strategies based on OBB vs. HBB fundamentally affect object detection performance and generalizability in real-world vineyard analytics [54,55].

To the best of our knowledge, no studies have specifically investigated viticultural object detection through an in-depth examination of object detectors within a single architectural lineage. Furthermore, there are no known test cases utilizing data sourced from three distinct geographical regions particularly Europe, South America, and Asia altogether. Likewise, there is a lack of experimental setups that incorporate substantial heterogeneity in grape bunches, encompassing variations in species, spectral ranges, image quality, acquisition protocols, and sensor types.

Hence, we have undertaken this study to undertakes a comprehensive comparison of the performances of various YOLO detectors particularly YOLOv8–YOLOv12, including OBB variant over multiple vineyard datasets spanning RGB and NIR domains under different acquisition modes, timings, geographies and protocols. The research is pivoted around testing the generalization abilities of these object detectors over different grape varieties with missing imaging protocols to examine how addressing these interconnected gaps will be essential for developing vineyard detection systems that are not only accurate but also robust, generalizable, and practically deployable in diverse viticultural environments. In this work, primarily there are three publically available datasets with one of them stratified into eight subdomains obtained by classifying as cultivar based on type of grape (red versus green), spectral modality (RGB versus NIR), and image-quality regime (normal versus poor). For example, "Green Grapes – RGB Normal" or "Red Grapes – NIR Poor". In this way, the dataset is broadly presented and formally defined in Section 2 and summarized in Table 1. This results in ten datasets which are described in Table 1.

Specifically, this study pursues four main aims:

A1. Benchmarks YOLOv8–YOLOv12 and an orientation-aware YOLOv8-OBB across three geographically distinct datasets (Italy, Brazil, China), spanning RGB and NIR spectra and both normal and degraded image quality regimes.

A2. Quantify detection reliability using conventional metrics (precision, recall, F1, mAP@0.5, mAP@0.5:0.95) alongside a FCR and variance across repeated runs, to capture both average performance and stability under domain shift.

A3. Assess the efficacy of OBB to check cluster detection failure linked to rotation, occlusion, and elongated bunch morphology in vineyard canopies, relative to conventional HBB, particularly in NIR and degraded imaging conditions.

A4. Derive deployment guidance that maps model behaviour to viticultural tasks such as yield estimation, precision spraying, and autonomous vineyard robotics so that practitioners can select architectures consistent with their accuracy, reliability, and latency requirements.

2. Materials and methods

This study consists of three major steps and begins with downloading publicly available grape image datasets from three sources which were curated in academic and commercial orchards in distinct seasons, timings of the day and illumination conditions across three different

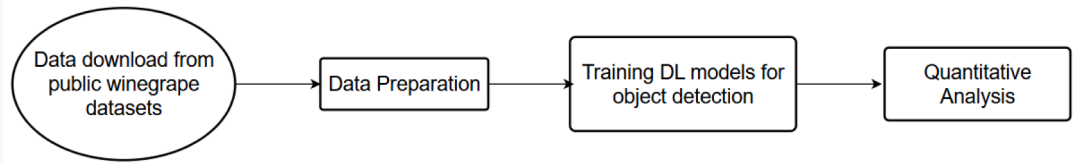
Table 1
Dataset characteristics.

	VINEyard Piacenza Image Collections - *VINEPICS*	Embrapa *WGISD*: Embrapa Wine Grape Instance Segmentation Dataset	Grape multimodal object detection and semantic segmentation dataset
Total images	238	300	7908
Curator	Universita Cattolica del Sacro Cuore, Piacenza, Italy	Brazilian Agricultural Research Organization (EMBRAPA), Brazil	Anhui Agricultural University, China
Nature of data	Instance Segmentation	Instance Segmentation	Semantic Segmentation
Image Resolution	1280×720	2048×1365	1280×720
Volume	345 MB	288 MB	39.08 GB
Sensor	D435 Intel Realsense camera	Canon EOS REBEL T3i DSLR camera and Motorola Z2 Play smartphone	Azure Kinect DK
Camera Angle	45° and 90°	90° at eye level	90°×59°/75°×65°
Orientation	Portrait	Landscape	Landscape
Spectra	RGB	RGB	RGB and Near Infra-Red
Grape Species	Red Globe (red), Cabernet Sauvignon (red), Ortrugo (white)	Chardonnay (white), Cabernet Franc (red), Cabernet Sauvignon (red), Sauvignon Blanc (white), Syrah (red)	Not specified
File Structure	Multidate images and organized into subdirectories and arranged periodically. Images collected in 2022 are named using timestamps.	WGISD -> images and annotations	Described in figure no 2
Source	https://doi.org/10.5281/zenodo.7866442	https://doi.org/10.5281/zenodo.3361736	https://www.scidb.cn/en/detail, ?dataSetId=84fa458dfc854fba8ce578b6d826b9c8

geographical locations (Fig. 1). These datasets cumulatively comprise of a heterogenous mix of winegrape images in terms of:

a) Geographies
b) Varying spatial resolutions owing to different sensor models
c) Spectra – RGB & NIR
d) Grape varieties
e) Image acquisition protocols
f) Dataset imbalances

All the datasets were already annotated to cater for instance segmentation by first hand curators. Within each of the ten domains, images

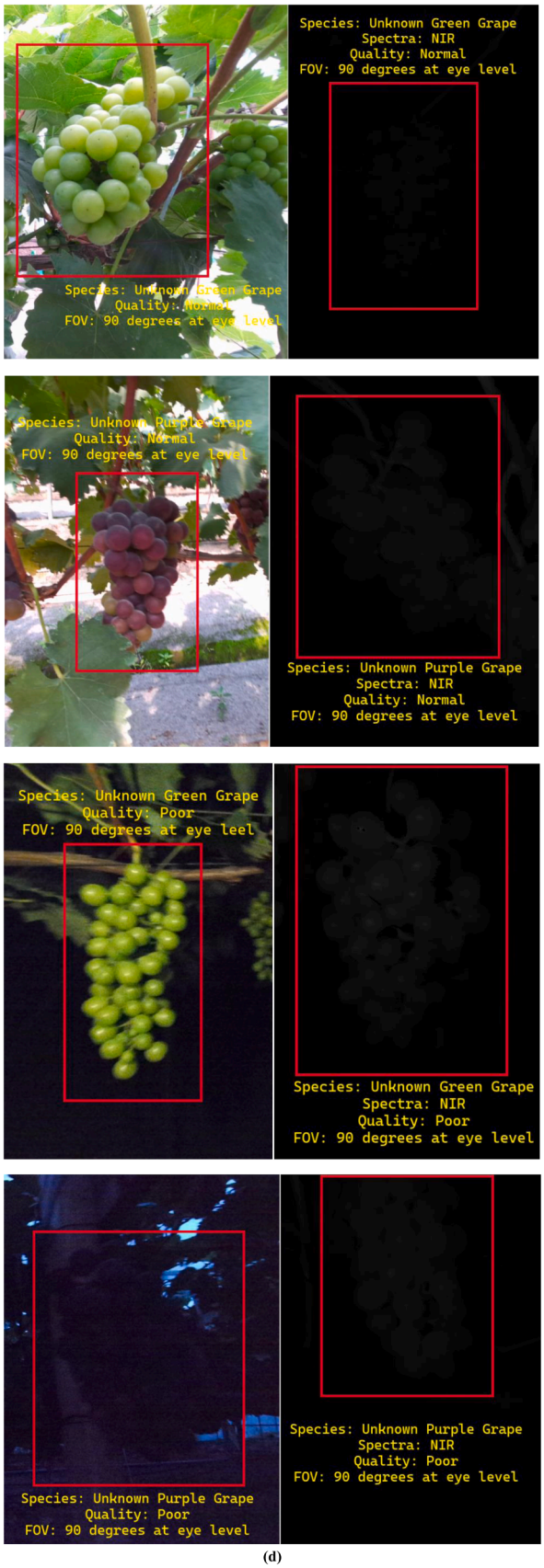


(a)

(b)

(c)

Fig. 1. (a): Overall workflow diagram used in this research; (b) example images of VINEPICS 2021 dataset with bunches highlighted in red anchor boxes, Universita Cattolica del Sacro Cuore (IT) (c) Example images of WGISD dataset with bunches highlighted in red anchor boxes, EMBRAPA Informatics institute (BR) (d) multimodal grapes dataset – object detection and semantic segmentation with bunches highlighted in red anchor boxes, anhui agricultural University.



(d)

Fig. 1. (*continued*).

were split into training, validation, and test subsets with a 70/20/10 ratio at the image level (Table 2). For the public datasets used here, explicit vine, row, or acquisition-sequence identifiers were not consistently available, so it was not possible to enforce strict vine-level

Table 2
Per-domain image and grape-cluster counts for training, validation and test splits.

Dataset	Cultivar	Train images (clusters)	Val images (clusters)	Test images (clusters)
Italy_VINEPICS 2021	Red/ white mix	166 (1582)	47 (348)	25 (473)
Brazil_WGISD	Red/ white mix	210 (3102)	60 (886)	30 (444)
China_Green_RGB_Normal	Green	667 (2551)	190 (729)	96 (365)
China Green_RGB_Poor	Green	667 (2551)	190 (729)	96 (365)
China Green_NIR_Normal	Green	700 (3375)	200 (964)	100 (483)
China Green_NIR_Poor	Green	700 (3375)	200 (964)	100 (483)
China Red_RGB_Normal	Red	710 (2863)	203 (818)	102 (410)
China Red_RGB_Poor	Red	710 (2863)	203 (818)	102 (410)
China Red_NIR_Normal	Red	690 (2981)	197 (851)	99 (427)
China Red_NIR_Poor	Red	690 (2981)	197 (851)	99 (427)

separation across splits. To reduce leakage, we visually inspected directories and removed near-duplicate frames of the same vine before splitting, but residual correlations between images from neighbouring vines or the same row cannot be fully excluded. All subsequent performance and generalization analyses therefore assume independence at the image level rather than at vine or block level. The training dataset was then used to train the six deep learning models and their performance in object detection was evaluated using the test dataset.

2.1. Study site and data preparation

This study was conducted on annotated images curated over three different grape orchards (Fig. 1b, 1c and 1d). The first dataset was curated in experimental vineyard located in Department of Sustainable Crop Production, Universita Cattolica del Sacro Cuore, located at Piacenza, Italy. The second dataset was curated in Guaspari Winery, located at Espírito Santo do Pinhal, São Paulo, Brazil by EMBRAPA Informatics Institute. The third dataset was curated by Anhul Agricultural University in its experimental vineyards. The other features of all three datasets are provided in Fig. 1 and Table 1. A consolidated overview of the ten evaluated domains, including the exact number of images and annotated grape clusters used for training, validation and testing, is provided in Table 2. Each domain corresponds to either a standalone dataset (VINEPICS, WGISD) or a stratified subdomain of the Chinese multimodal dataset, defined by cultivar (red/green), spectral band (RGB/NIR) and quality regime (normal/poor). (Fig. 2)

2.2. Objective quantification of image quality regimes

The Chinese multimodal dataset distinguishes between “normal” and “poor” image quality regimes based on the original curator’s annotations. To validate and quantify this distinction, we retrospectively computed two reference-free image quality metrics for all RGB and NIR images:

(i) The BRISQUE score, which captures natural-scene statistics deviations (higher values indicate lower perceived quality), and
(ii) A simple blur index, defined as the variance of the Laplacian of the grayscale image (lower values indicate stronger blur).

All metrics were computed on the original-resolution images prior to resizing to 640 × 640. For each modality and quality regime, we have reported the distribution of BRISQUE and blur scores (median, interquartile range, mean ± standard deviation) in Fig. 10, and we used a non-parametric Mann - Whitney U test to assess whether “normal” and “poor” subsets differ significantly.

2.3. Implementation of object detection models

All datasets were resized to 640 × 640 pixels and split into training, validation, and test subsets with a 70/20/10 % ratio at the image level, yielding comparable proportions of grape-cluster instances across splits. All YOLO variants (YOLOv8x, YOLOv8x-OBB, YOLOv9e, YOLOv10x, YOLOv11x, and YOLOv12x) were trained on their largest “x” configurations under a single, standardised optimisation protocol to ensure a controlled and reproducible comparison. Concretely, we used stochastic

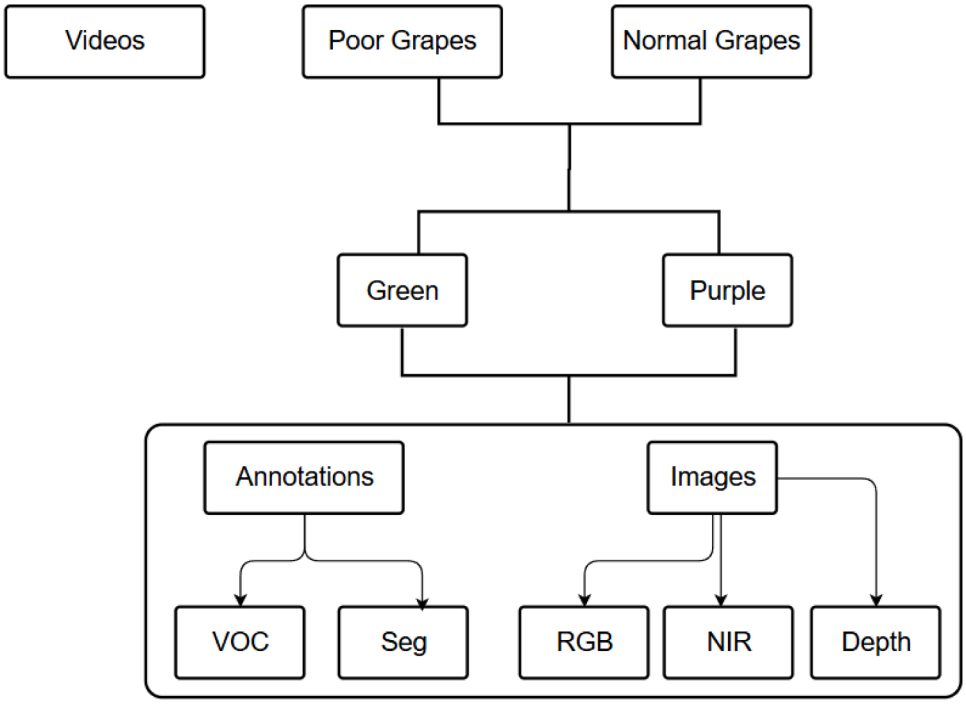


Fig. 2. Data structure of grape multimodal object detection and semantic segmentation.

gradient descent (SGD) with an initial learning rate of 0.001, a cosine decay schedule down to 0.0001, momentum = 0.946, weight decay = 0.0005, batch size = 16, and a maximum of 100 epochs with early stopping based on validation performance (training was stopped once validation accuracy plateaued or began to decline). During the first two epochs we applied a warm-up strategy with reduced momentum (0.8) and an increased bias learning rate (0.1) to stabilise optimisation. After training, all models were exported to TorchScript for streamlined evaluation in terms of precision, recall, mAP, and FCR. We deliberately did not perform architecture-specific hyperparameter searches like per-model tuning of learning rate, weight decay, or augmentation strength and instead treat each detector as an "off-the-shelf" model that a vineyard practitioner or robotics engineer could realistically deploy under a fixed training budget; this ensures identical training conditions across architectures but also implies that some models are likely not operating at their individually optimal hyperparameter configuration.

In addition to the HBB variants, we included YOLOv8x-OBB, an orientation-aware detector implemented in the Ultralytics framework. YOLOv8x-OBB was initialised from the publicly available yolov8x-obb.pt checkpoint, pre-trained on the DOTA v1.0 aerial oriented object dataset [54], and all layers (backbone, neck, and OBB detection head) were unfrozen and jointly fine-tuned using the optimisation protocol described above. All horizontal YOLO variants (YOLOv8x, YOLOv9e, YOLOv10x, YOLOv11x, YOLOv12x) were initialised from the official Ultralytics COCO-pretrained weights (e.g. yolov8x.pt, yolov9e.pt, yolov10x.pt). Thus, all models benefit from large-scale pre-training, but YOLOv8x-OBB is the only architecture whose pre-training explicitly targets oriented objects, a difference we revisit in the Discussion when interpreting its advantage in rotated and NIR-poor scenes.

3. Data augmentation

To improve robustness and avoid overfitting, all models were trained with the same on-the-fly data augmentation policy applied only to the training split while validation and test images were left unaltered. Geometric augmentations included random horizontal flipping with a probability of 0.5, random vertical flipping with a probability of 0.2, and random in-plane rotation sampled uniformly in the range with a probability of 0.5. We additionally used random scaling in the range 0.8–1.2 and random translation up to 10 % of the image size to simulate viewpoint changes. Mosaic augmentation was enabled with a probability of 0.5 during the first 80 % of training epochs and disabled in the final 20 % to stabilise convergence. Photometric augmentations consisted of random brightness and contrast adjustments in the range ±20 %, HSV saturation shifts of ±15 %, and Gaussian blur with $\sigma \in [0,1]$applied with a probability of 0.3. The same augmentation settings (types, parameter ranges, and application probabilities) were used for all YOLO variants and for all three datasets, so that no model or dataset benefited from a differential augmentation scheme.

3.1. Performance evaluation

To assess the object detection performance of the various YOLO models, six evaluation metrics were employed: precision, recall, F1-score, mAP@0.5, mAP@0.5:0.95, and FCR. For each model-dataset pair, detections were matched to ground-truth boxes at a fixed Intersection over Union (IoU) threshold of 0.5 and unless otherwise mentioned, a fixed confidence threshold of 0.25 was used with no additional score calibration. True positives (TP), false positives (FP), and false negatives (FN) were first accumulated over all test images in a given domain, and precision, recall, F1-score, and FCR were then computed from these pooled counts. Precision measures the proportion of correctly identified positive instances relative to the total number of predicted positives (Eq. (1)), whereas recall (Eq. (2)) represents the proportion of correctly identified positive instances out of all actual positive samples. The F1-score (Eq. (3)) is the harmonic mean of precision and recall and summarises their balance. The mAP@0.5, calculated as the mean of the average precision across categories at IoU = 0.5, serves as a key indicator of the model's detection accuracy, while mAP@0.5:0.95 extends this evaluation by averaging the AP over multiple IoU thresholds from 0.50 to 0.95 in steps of 0.05, providing a more comprehensive assessment of robustness. Finally, the FCR is an error-centric reliability metric that quantifies the proportion of misclassified instances among all detected or missed positives. FCR × 100 % can therefore be interpreted as the percentage of grape clusters (detected or missed) that are wrongly classified. Unlike the F1-score, which combines precision and recall via their harmonic mean, FCR directly measures the concentration of failures among all positive instances and is particularly suited for assessing operational risk in robotic deployments. In the Results (Sections 3.3 and 3.9), we show that models with similar F1 can exhibit substantially different FCR, highlighting the additional information that FCR provides beyond standard accuracy metrics. These metrics are calculated as follows:

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{1}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{2}$$

$$\boldsymbol{mAP} = \frac{1}{\boldsymbol{K}} \sum_{i=0}^{k} (\boldsymbol{AP})\boldsymbol{i} \tag{3}$$

$$\boldsymbol{FCR} = \frac{(\boldsymbol{FP} + \boldsymbol{FN})}{(\boldsymbol{TP} + \boldsymbol{FP} + \boldsymbol{FN})} = 1 - \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FP} + \boldsymbol{FN})} \tag{4}$$

Here, TP, FP, and FN denote the numbers of true positive, false positive, and false negative object instances, respectively. The variable k represents the total number of object classes, while $(AP)i$ denotes the average precision computed for the i^{th} class among these k classes. The AP corresponds to the area under the precision–recall curve for a given class, providing a quantitative measure of detection accuracy across varying confidence thresholds.

For evaluation of the OBB variant, the predictions and labels were converted to their minimum enclosed HBB, and standard IoU (axis-aligned) was used to compute mAP@0.5, mAP@0.5:0.95, precision, recall, F1, and FCR. This ensured that all models (HBB and OBB) were evaluated under an identical metric definition. This is mainly done because axis-aligned IoU tends to underestimate the localization advantage of OBB in highly rotated or elongated clusters, so the reported gains for YOLOv8-OBB are conservative.

4. Results

4.1. Comparison of best performing models across all datasets

Across ten vineyard datasets, no single detector dominates. In clean RGB, YOLOv11–YOLOv12 lead whereas under mixed quality and modality whereas YOLOv10 stays consistently near the top (Fig. 3(a-c)). In NIR and scenes with challenging contrasts, YOLOv8-OBB is strongest (Fig. 3(d)). YOLOv9 occasionally ranks first but varies more across domains. Full best-per-dataset results are in Table 3.

4.2. Multi-domain generalizability

Cross-domain behaviour favours geometry or scale-robust models: YOLOv10 and YOLOv8-OBB generalize best across RGB/NIR and quality shifts. YOLOv11 is very strong in RGB; YOLOv12 is high in clean RGB but brittle under shift; YOLOv9 is recall-leaning with FP spikes in NIR. See Table 4 (qualitative summary).

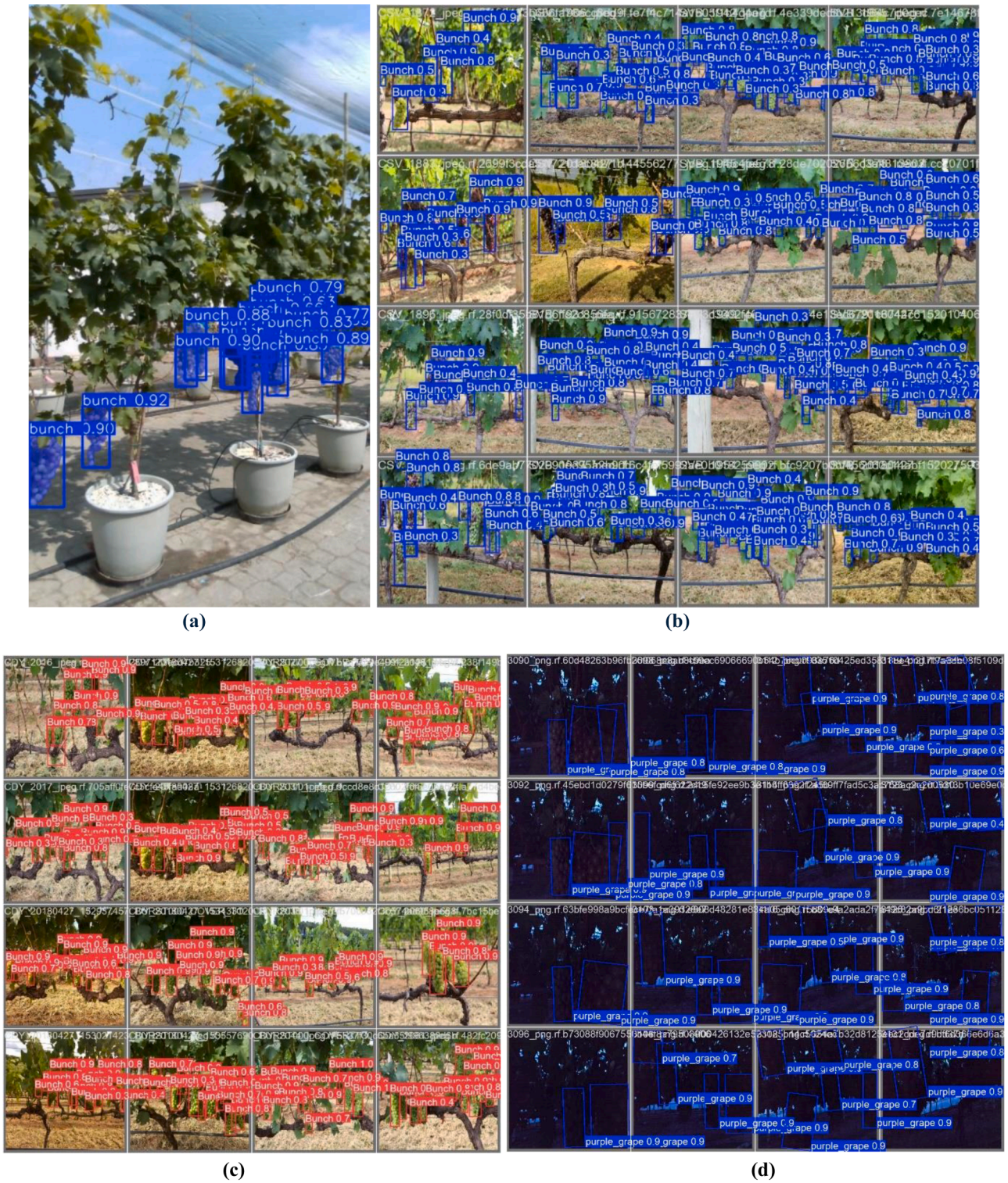


Fig. 3. (a) Predictions on VINEPICs 2021 by YOLOv11 (b) predictions on WGISD EMBRAPA by YOLOv12 (c) predictions on WGISD EMBRAPA by YOLOv10 (d) Predictions on Red Grape NIR Poor by YOLOv10.

4.3. Orientation-aware performance gains

The integration of orientation awareness through OBBs in models such as YOLOv8-OBB plays a critical role in improving detection performance in complex agricultural scenarios, especially where target objects like grape clusters exhibit irregular, rotated, or elongated shapes. This architectural enhancement enables the model to fit bounding boxes that align with the actual orientation of the object, rather than using axis-aligned boxes that often enclose significant background pixels. Below is a comprehensive analysis of the orientation-aware performance gains observed across the datasets: Orientation-aware regression (YOLOv8-OBB) improved localization and background rejection in non-axis-aligned clusters. Across WGISD and VINEPICS, mAP@0.5 and mAP@0.5:0.95 increased by approximately 0.03 and 0.05 respectively with respect to YOLOv8 (Fig. 4(a)). In the NIR domain (Fig. 4(b)), these gains manifested as both higher F1 and lower FCR. On the most

Table 3
Best performing object detection models across different datasets.

Spectrum	Dataset		Best Model	F1-Score	mAP@0.5	mAP@0.5:0.95	FCR
RGB	**VINEPICS 2021 (Italy)**		**YOLOv11**	**0.864**	**0.91**	0.48	0.185
	WGISD EMBRAPA (Brazil)		**YOLOv8-OBB**	**0.865**	**0.9**	0.7	0.29
RGB & NIR	**Grape multimodal object detection and semantic segmentation dataset (China)**	Green Grapes - RGB Poor	YOLOv9	0.884	0.9	0.76	0.176
		Green Grapes - RGB Normal	**YOLOv11/ v12**	**0.91**	**0.954**	**0.8**	**0.095**
		Green Grapes - NIR Poor	**YOLOv10**	**0.877**	**0.88**	0.72	**0.152**
		Green Grapes - NIR Normal	YOLOv12	0.82	0.86	0.67	0.39
		Red Grapes - RGB Poor	YOLOv11	0.87	0.92	0.5	0.18
		Red Grapes - RGB Normal	**YOLOv10**	**0.86**	**0.93**	**0.79**	**0.14**
		Red Grapes - NIR Normal	YOLOv9	0.726	0.77	0.512	0.274
		Red Grapes - NIR Poor	**YOLOv8-OBB**	**0.911**	**0.94**	**0.76**	**0.086**

challenging *Red Grapes – NIR Poor* subset, F1 increased from 0.888 ± 0.008 with the strongest horizontal-baseline YOLOv10 to 0.911 ± 0.007 with YOLOv8-OBB, while FCR decreased from 0.110 ± 0.009 to 0.086 ± 0.008 (absolute change −0.024, ≈ 22 % relative). Given that FCR = (FP + FN)/(TP + FP + FN), this corresponds to a reduction in the misclassification ratio from 11.0 % to 8.6 % of all positive instances. Similar, slightly smaller improvements in FCR were observed on Green Grapes – NIR Normal and NIR Poor. Overall, YOLOv8-OBB handles rotated and elongated clusters, reduces leaf confusions, and the gains are largest in NIR-normal and poor-quality images.

4.4. Effect of image quality on model performance

Variations in image quality such as blur, lighting imbalance, occlusion, or spectral noise significantly impacted detection performance across all YOLO variants. These degradations reduce edge clarity, suppress grape-texture contrast, and increase misclassification risk. However, not all models responded equally.

Fig. 5(a) illustrates the distribution of F1-score drop under degraded imaging. YOLOv12 showed the highest variability, with some extreme drops, while YOLOv11 and YOLOv10 exhibited tighter distributions, highlighting stronger resilience. YOLOv8-OBB stood out for its low median F1-drop and narrow spread, thereby confirming its robustness in noisy or rotated scenes due to orientation-awareness.

Fig. 5(b) presents mAP@0.5 drop trends across domains. All models suffered most in *Red Grapes – NIR Poor dataset*, confirming this as the most challenging domain due to its low chromaticity and structural ambiguity. While YOLOv8-OBB showed the least degradation in Green NIR, YOLOv12 exhibited unusual performance boosts in *Red Grapes - NIR Poor*, likely due to architecture-specific feature matching rather than general robustness.

Model-wise, it was observed that:

- **YOLOv8**: Strong in clean RGB but degraded notably under NIR and poor-quality lighting.
- **YOLOv8-OBB**: Showed minimal performance drop in both RGB and NIR, especially for rotated or occluded clusters, making it optimal for real-world variability.
- **YOLOv9**: Balanced precision-recall in RGB but unstable in NIR, where over-detection inflated false positives.
- **YOLOv10**: Most consistent across all conditions, with modest drop in both F1 and mAP, marking it a strong general-purpose model.
- **YOLOv11**: High average scores but reduced robustness under quality degradation, with elevated prediction variance.
- **YOLOv12**: Inconsistent but performed well in specific NIR-poor domains and yet lacked cross-domain reliability.

4.5. Precision-recall tradeoffs

The precision-recall trade-off is a fundamental concept in evaluating the performance of object detection models. Precision measures the accuracy of positive predictions, i.e., how many detected grape clusters were correct, while recall assesses the model's ability to identify all relevant instances, like how many true grape clusters were found. Practically, improving one often comes at the cost of the other. High precision can be achieved by being conservative in detections, by risking more false negatives whereas high recall involves capturing more true objects but may introduce more false positives. This trade off in grape bunch detection due to under-detection (low recall) leads to inaccurate yield estimation, and over-detection (low precision) leads to resource misallocation or overestimation.

4.5.1. Precision-optimized models

Models emphasizing precision favour conservative detections that minimize false positives but may sacrifice recall. Among all variants, YOLOv12 exhibited the strongest precision-oriented behaviour, consistently delivering the lowest FCR across clean RGB and NIR-poor datasets. This precision bias stems from its selective confidence thresholds and tighter bounding-box filtering, making it highly suitable for vineyard operations where false positives incur tangible costs such as targeted pesticide spraying, disease hotspot identification, or robotic fruit picking where actuator errors must be minimized. YOLOv8-OBB achieved similarly high precision yet maintained stronger recall stability, owing to its orientation-aware bounding boxes that preserve geometric fidelity and reduce background confusion. This combination of accuracy and spatial awareness makes YOLOv8-OBB particularly effective for tasks demanding both structural consistency and positional accuracy, including harvest automation, vineyard zoning, and canopy-structure phenotyping.

4.5.2. Recall-optimized models

Models emphasizing recall prioritize detection completeness, ensuring that nearly all grape clusters are identified even under challenging visual conditions such as blur, occlusion, or low contrast. YOLOv9 demonstrated the most aggressive recall behaviour across both RGB and NIR domains, adopting a liberal confidence threshold that maximizes true-positive detections but inevitably introduces more false alarms, reflected in its higher FCR. This trade-off makes YOLOv9

Table 4
Performance of different object detection models in cross-domain generalization.

Model	Strengths	Cross domain evaluation	Insights
YOLOv11	Top performing in high-quality RGB datasets; competitive in poor quality RGB and moderate quality NIR datasets.	Trained: **Green Grapes – RGB Normal**: F1 = 0.91, mAP@0.5 = 0.954 Validated: **Green Grapes – RGB Poor**: F1 = 0.87 **Green Grapes – NIR Poor**: F1 = 0.854, mAP@0.5 = 0.882	C3K2 modules and depth wise convolutional layers improved spatial localization and robustness to lighting variations.
YOLOv10	Excellent generalization across RGB and NIR; resilient in poor-quality and low-contrast conditions.	Trained: **Green Grapes – NIR Poor**: F1 = 0.877, mAP@0.5 = 0.88, FCR = 0.152 (lowest) Validated: **Red Grapes – RGB Normal**: F1 = 0.86, mAP@0.5 = 0.93	NMS-free dual-label assignment enhanced the robustness in noisy, blurred, and low-contrast images.
YOLOv8-OBB	Strongest in structurally complex, NIR-heavy, and occluded domains due to use of OBB.	Trained: **Red Grapes – NIR Poor**: F1 = 0.911, mAP@0.5 = 0.94, FCR = 0.086 (lowest) Validated: **Green Grapes – NIR Normal**: F1 = **0.81** **WGISD EMBRAPA**: F1 = 0.865, mAP@0.5 = 0.90	Orientation-aware detection helped in localizing rotated clusters and improved discrimination from the background.
YOLOv12	High performance in clean RGB domains, but reduced robustness under noise, spectral variation, and NIR imaging.	Trained: **Green Grapes – RGB Normal**: F1 = 0.91, mAP@0.5 = 0.954 Validated: **Green Grapes – RGB Poor**: Recall = 0.79 **Red Grapes – NIR Normal**: Recall = 0.57, F1 = 0.726	Attention modules are optimized for clean texture and shape features but struggled with degraded NIR or noisy inputs.
YOLOv8	Reliable in normal RGB conditions but shows low adaptability across NIR domains and poor-quality scenarios.	Trained: **Green Grapes – RGB Normal**: F1 = 0.90 Validated: **Red Grapes – NIR Normal**: F1 = 0.69, mAP@0.5:0.95 = 0.512	Anchor-free heads and simplified architectural elements reduced generalization to NIR.
YOLOv9	High-recall model with strong performance in RGB Poor and NIR Normal domains but suffers from inconsistency and elevated false positives in noisy NIR scenes.	Trained: **Green Grapes – RGB Poor:** F1 = 0.88, mAP@0.5 = 0.90, FCR = 0.176 Validated: **Red Grapes – NIR Normal:** F1 = 0.726, Recall = 0.645 **Red Grapes – NIR Poor:** F1 = 0.91, but high FP and FCR = 0.2737	Prioritized recall across domains, often detecting small or occluded clusters, but susceptible to over-detection and false positives in low-contrast or cluttered NIR scenes.

especially valuable in applications where under-detection is more detrimental than over-counting such as yields estimation, phenotyping, or canopy-completeness mapping, where comprehensive coverage outweighs precision. YOLOv11, while also achieving high recall, maintains better balance with precision, offering a more conservative option for general vineyard monitoring or semi-automated deployments where both completeness and stability are required.

4.5.3. Balanced precision-recall models

Detectors maintaining similar precision and recall achieve the highest F1-scores and overall reliability. YOLOv10 exhibited the most even balance across spectral and quality domains, sustaining performance in both RGB and NIR conditions. YOLOv11 followed closely with a mild recall bias but minimal variance, while YOLOv8-OBB preserved this balance under geometric complexity thanks to its orientation-aware design. These balanced models provide dependable results for general vineyard monitoring, harvest robotics, and phenotyping where both false alarms and omissions must be minimized. (See Table 5).

4.6. Quantitative assessment of trade-offs

Across all datasets, models achieving precision and recall within 5–7 % of each other consistently reached the highest F1-scores, confirming that balance rather than peak values determines reliability in viticultural imaging. Averaged over normal and degraded imagery, performance dropped by approximately $\Delta F1 = -0.03$, $\Delta mAP@0.5 = -0.03$, and $\Delta mAP@0.5{:}0.95 = -0.07$, while the FCR increased by about 0.05. These shifts quantify the cost of spectral noise, glare, and motion blur on detector stability.

Among all variants, YOLOv10, YOLOv11, and YOLOv8-OBB maintained the most stable precision–recall equilibrium across RGB and NIR domains. YOLOv10 preserved recall with minimal precision loss under degraded conditions, YOLOv11 retained low variance in both metrics, and YOLOv8-OBB sustained high F1-scores in NIR-poor scenes owing to its orientation-aware localization. Together, these models demonstrated resilience to quality degradation and cross-spectral variability, a prerequisite for autonomous vineyard operations where lighting and canopy structure fluctuate dynamically. Representative quantitative examples are summarized in Table 5.

4.7. Generalization in complex contexts

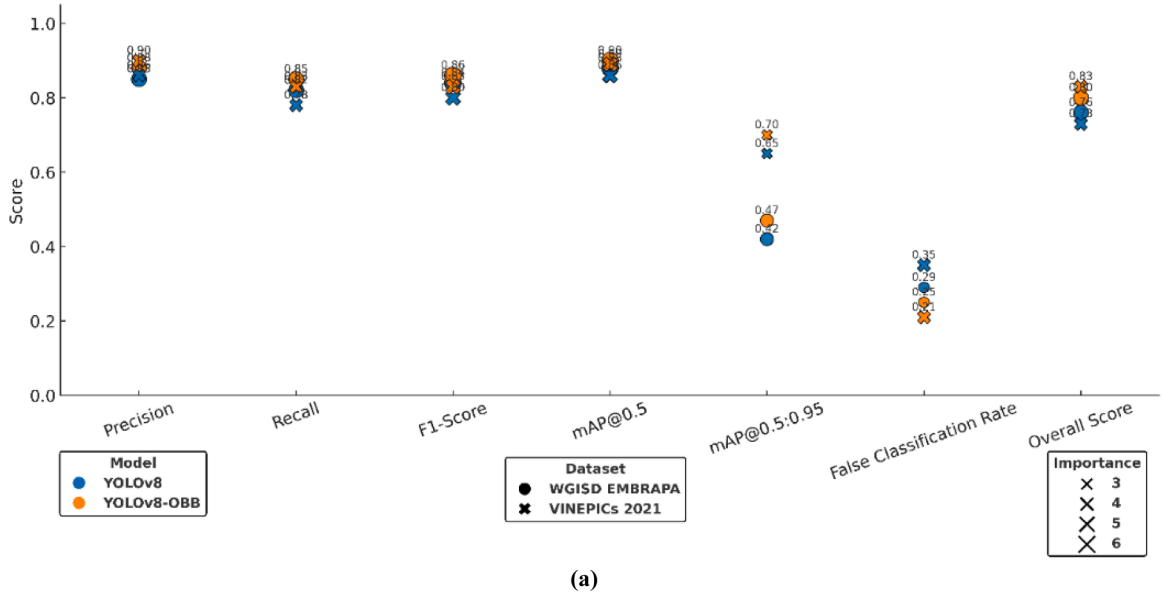
Generalization reflects a model's capacity to sustain detection accuracy across grape varieties, spectral modalities, and image degradations typical of vineyard environments. Models emphasizing geometric and multi-scale cues generalized best. YOLOv10 and YOLOv11 consistently transferred structural knowledge between red and green grape datasets, maintaining high F1-scores despite differences in cluster compactness and leaf density. Their balanced feature aggregation captured shared spatial patterns, grape-leaf boundaries and cluster regularity independent of colour.

Across spectral domains, where texture weakens and colour gradients collapse, YOLOv8-OBB and YOLOv10 bridged the RGB–NIR gap most effectively. Orientation-aware bounding boxes and scale-adaptive heads allowed them to rely on shape and spatial arrangement rather than chromatic contrast, yielding domain-invariant representations valuable for multispectral sensors.

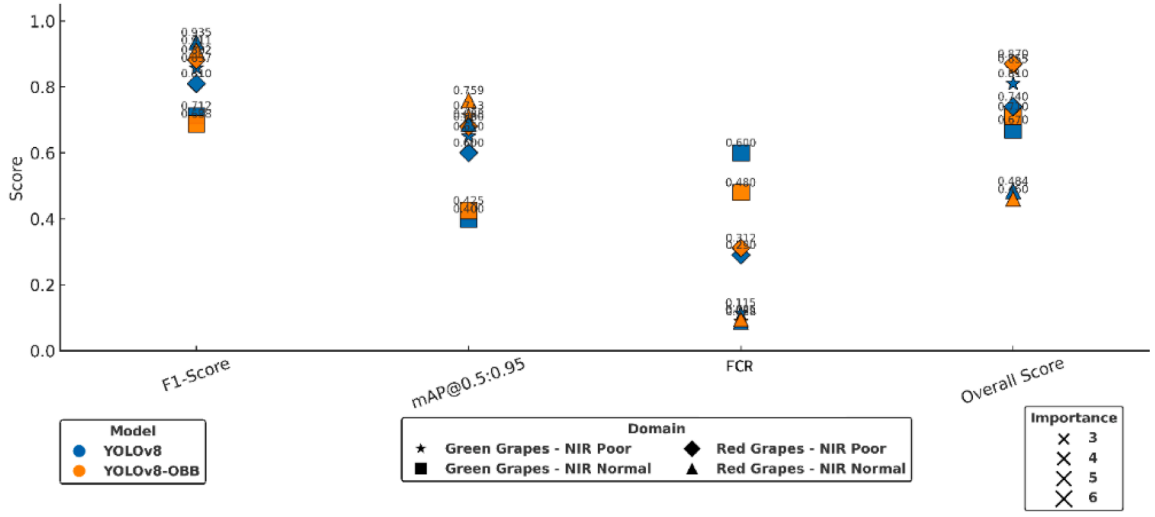
Under environmental variability - occlusion, glare, and motion blur, YOLOv11 remained strong in cluttered RGB scenes, while YOLOv8-OBB excelled under heavy shadow and irregular geometry. Even in NIR-poor imagery, both YOLOv8-OBB and YOLOv10 sustained $F1 > 0.85$, whereas YOLOv9 and YOLOv12 degraded sharply, revealing susceptibility to low-contrast inputs. When transferred across datasets, only YOLOv8-OBB preserved high detection reliability, confirming its orientation-aware encoding as the most portable across vineyard domains. Table 5 summarizes comparative strengths across domain, spectral, and occlusion dimensions. While YOLOv11 remains robust under occlusion and clutter in-domain RGB imagery, its cross-dataset transfer is more limited, justifying its classification as high in occlusion handling but only moderate in cross-domain generalization in Table 6.

4.8. Statistical spread of metrics

The standard deviation of detection metrics across datasets



(a)



(b)

Fig. 4. (a) Performance evaluation of YOLOv8 and YOLOv8-OBB on WGISD EMBRAPA and VINEPICs 2021 datasets (b) Performance Evaluation of YOLOv8 and YOLOv8-OBB on NIR domain datasets.

quantifies each model's stability and generalization potential (Fig. 6). A low spread indicates consistent behaviour under domain shift, which is essential for autonomous operation in dynamic vineyard environments.

Among all models, YOLOv11 exhibited the lowest variance across F1-score, precision, recall, mAP@0.5, and FCR, confirming it as the most statistically stable architecture. YOLOv10 followed closely, with minimal dispersion in precision and mAP, reflecting strong general-purpose reliability. YOLOv8-OBB showed slightly higher overall variance but remained exceptionally consistent in NIR datasets, where its orientation-aware detection mitigated spectral and geometric distortions.

Contrastingly, YOLOv12 and YOLOv9 recorded the largest standard deviations particularly in recall and mAP signifying greater sensitivity to noise, lighting, and spectral shifts. Across domains, the mean mAP@0.5 variance was lower for RGB ($\approx$ 0.0065) than for NIR ($\approx$ 0.012), underscoring the stronger stability of colour imagery and the spectral fragility of NIR detection.

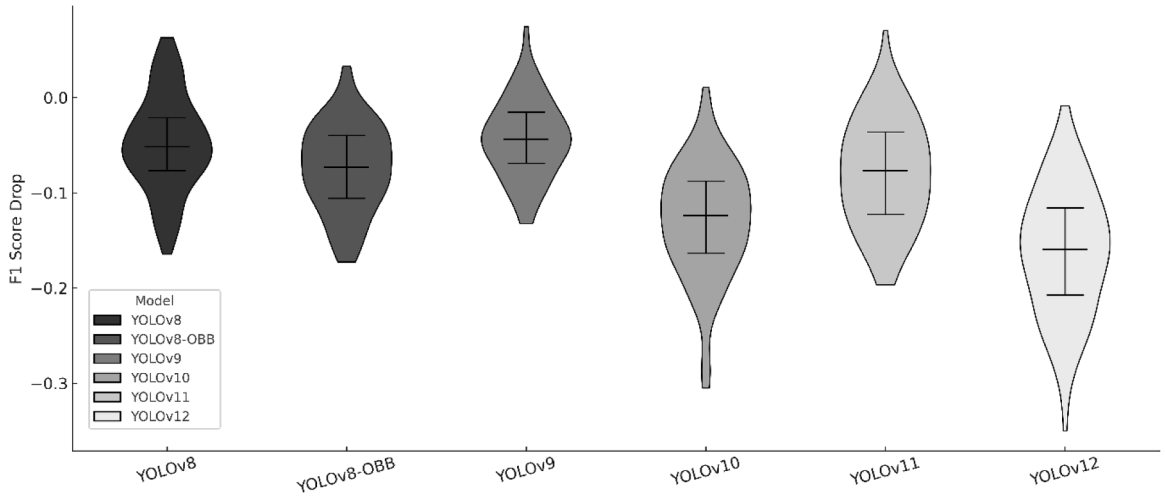
Overall, the statistical spread analysis corroborates earlier findings that YOLOv11 delivers the most predictable performance across conditions, while YOLOv10 and YOLOv8-OBB balance robustness and domain adaptability, making them the most reliable options for field-deployable vineyard vision systems.

4.9. Comprehensive error analysis

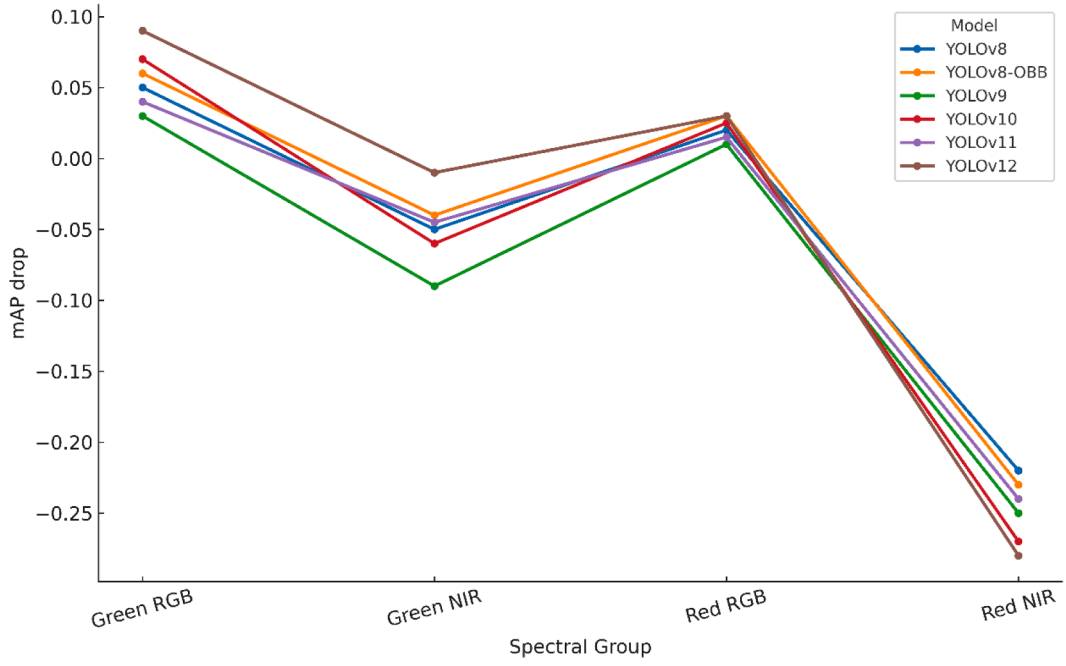
A detailed examination of FP, FN, and the combined FCR reveals the characteristic failure modes of each YOLO variant under field conditions (Fig. 7–9). Rather than absolute counts, these patterns highlight how spectral noise, geometry, and confidence calibration interact with architectural design choices.

4.9.1. Error dynamics

FP errors mainly arise from spectral ambiguities like intra vine leaf occlusions and background glare in NIR or structural confusion from tendrils in cluttered RGB scenes. YOLOv9, tuned for high recall, tends to



(a)



(b)

Fig. 5. (a) Distribution of F1-score drop across YOLO model variants (YOLOv8 to YOLOv12, including YOLOv8-OBB) under degraded image conditions (b) Comparison of mAP@0.5 drop across spectral domains and grape types for YOLO models.

over-detect, while YOLOv12, optimized for precision, suppresses FP at the cost of missing faint or occluded clusters. YOLOv10 maintains the most balanced behaviour through multi-scale feature fusion and robust objectness calibration. The orientation-aware YOLOv8-OBB yields the lowest FP rates in NIR-poor and oblique-angle datasets by aligning bounding boxes to true object geometry, though its recall drops slightly in extremely low-contrast RGB imagery.

4.9.2. *False-negatives*

Most FN instances stem from heavy occlusion, dense canopies, or low illumination. YOLOv10 and YOLOv11 preserve recall by leveraging contextual aggregation, while YOLOv12 and YOLOv9 underperform when contrast diminishes or targets overlap. These results reinforce that models emphasizing spatial context outperform those relying solely on chromatic texture (Fig. 7).

4.9.3. *Reliability via FCR*

The FCR integrates FP and FN to express real-world reliability at a fixed IoU and confidence threshold. FCR $\times$ 100 % is the percentage of positive instances (TP + FP + FN) that are misclassified as (FP + FN). The lowest FCRs ($\approx$ 8–10 %, 0.086 $\pm$ 0.008 for YOLOv8-OBB on *Red Grapes – NIR Poor*) occur for YOLOv8-OBB in NIR-poor domains whereas the highest ($\approx$ 35–40 %) appear for YOLOv12 under spectral distortion. These results demonstrate that geometric awareness and balanced thresholds, not extreme precision or recall bias, underpin dependable vineyard detection (Fig. 8). Overall, YOLOv10, YOLOv11, and YOLOv8-OBB exhibit the most controlled and operationally robust error profiles for robotic deployment.

Across all model and dataset combinations, F1 and FCR followed the expected inverse trend. The models with higher F1 generally exhibited a lower FCR, however, the two metrics were not redundant. Several cases

Table 5
Precision-recall balanced YOLO variants: metrics and underlying stability factors.

Model and Dataset	Precision	Recall	F1-Score	Reason of balance
YOLOv10 Red Grapes – RGB Normal	0.87–0.918	0.84–0.85	0.860–0.877	Multi-scale and stable across domains. 2 % difference between precision and recall; F1-score confirms well-balanced detection.
YOLOv11 Red Grapes – NIR Poor	0.90–0.916	0.83–0.90	0.864–0.908	Despite operating in a challenging NIR domain, this model maintained excellent balance, with high precision and minimal compromise on recall
YOLOv8-OBB WGISD EMBRAPA	0.88–0.959	0.85–0.87	0.865–0.911	Orientation-aware and domain-stable. A reliable model across different datasets with consistent metric alignment, reflecting adaptability

showed near-identical F1-scores with markedly different FCR values. Especially on the *Red Grapes – NIR Poor* subset, YOLOv9 and YOLOv8-OBB both reached F1 ≈ 0.91, yet YOLOv9 incurred an FCR of 0.2737 while YOLOv8-OBB remained at 0.086. Practically, this means YOLOv9 made roughly three times as many misclassified decisions per grape cluster at the chosen operating point, despite having a similar harmonic mean of precision and recall. This pattern reoccurred in other domains where differences in F1 were small (< 0.02) but FCR diverged, confirming that F1 alone can mask concentrated failure modes which were operationally critical for autonomous vineyard robots, whereas FCR directly exposed the density of costly mistakes.

Across three random seeds, all three detectors show tight confidence intervals, but YOLOv11 generally exhibits the smallest dispersion in both F1 and FCR, with typical 95 % CI half-widths around 0.005–0.007 (Fig. 9). On the Red Grapes–NIR Poor subset, YOLOv8-OBB achieved the highest F1 and lowest FCR among the three robust detectors ($F1 = 0.911 \pm 0.007$ vs. 0.888 ± 0.008 for YOLOv10 and 0.882 ± 0.009 for YOLOv11; $FCR = 0.086 \pm 0.008$ vs. 0.110 ± 0.009 and 0.118 ± 0.010). A paired bootstrap test on per-image F1 confirmed that YOLOv8-OBB significantly outperformed YOLOv10, with a 95 % confidence interval for ΔF1 of [0.014, 0.032] and $p < 0.05$. For FCR, the corresponding interval for ΔFCR was [−0.034, −0.018], and a Wilcoxon signed-rank test on per-image FCR also indicated a significant improvement ($p < 0.05$).

4.10. Objective characterization of normal vs. poor quality regimes

Fig. 10 summarizes the distribution of BRISQUE and blur scores for the "normal" and "poor" subsets of the Chinese dataset. In RGB, "normal" images exhibited substantially lower BRISQUE values than "poor" ones (median 23.0 vs. 33.0; mean ± SD 23.5 ± 6.5 vs. 34.0 ± 8.0; $p < 0.001$, Mann–Whitney U), indicating fewer statistical distortions. Conversely, the blur index was higher for normal images (median 250; mean ± SD 255 ± 85) than for poor images (median 150; 155 ± 70; $p <$

Table 6
Comparative synthesis of YOLO model generalization performance across spectral, structural, and dataset domains.

Model	Domain Robustness	Spectral Transfer	Occlusion Handling	Cross-Dataset Transfer	Overall Generalization
YOLOv10	High	High	Moderate–High	High	Very Strong
YOLOv11	High	Moderate	High	Moderate	Strong
YOLOv8-OBB	Moderate–High	Very High	High	Very High	Very Strong
YOLOv9	Moderate	Moderate	Low–Moderate	Low	Moderate
YOLOv12	Low–Moderate	Low	Moderate	Low	Weak

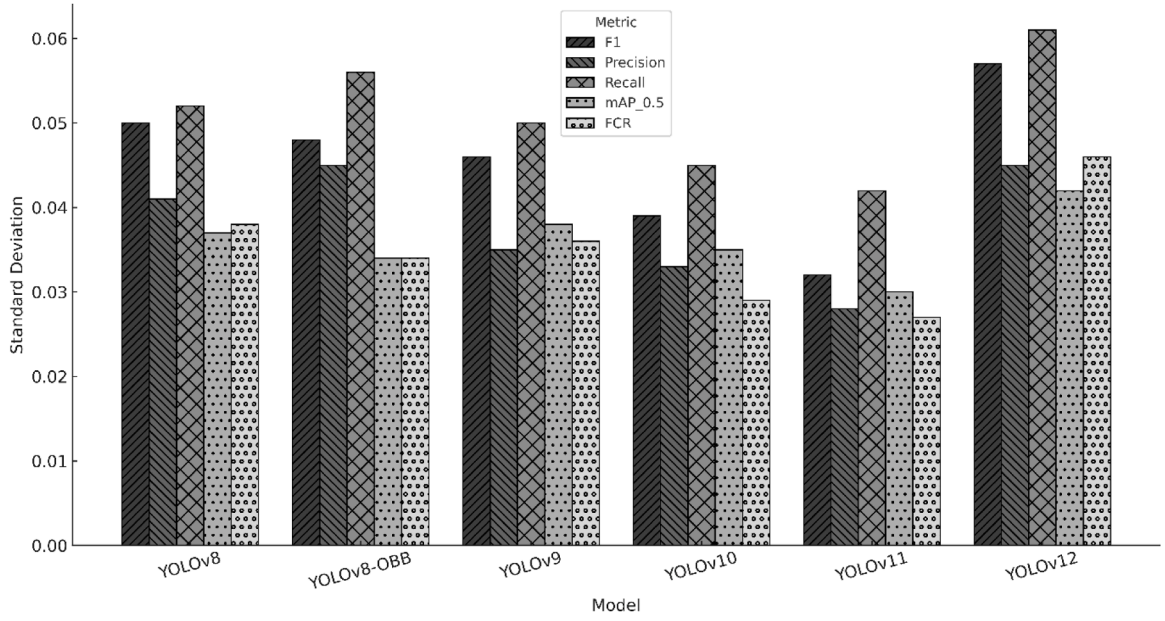


Fig. 6. Comparative standard deviation of key performance metrics.

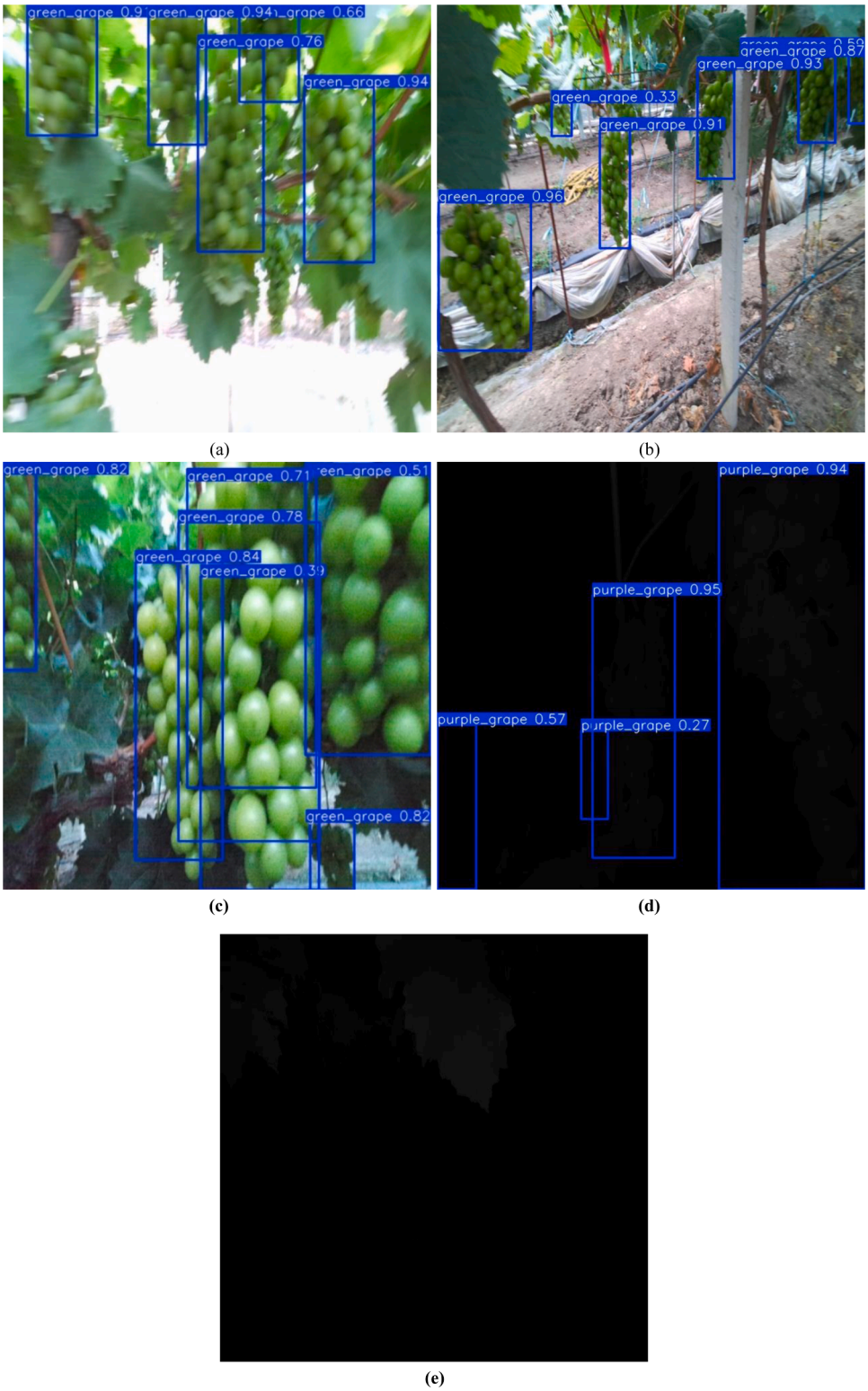


Fig. 7. (a) FN due to overexposure and blur (b) FN due to occlusion (c) FPs due to blur, heterogeneity and closeness of bunches (d, e) FPs due to underexposure and missed detection in NIR.

0.001), confirming that the "poor" regime systematically contains more motion blur and defocus. Similar trends were observed in NIR, with "normal" images again showing lower BRISQUE values than "poor" ones (median 27.0 vs. 43.0; mean ± SD 27.5 ± 7.5 vs. 44.0 ± 9.5; $p < 0.001$) and higher blur indices (median 220; 225 ± 80 vs. 115; 120 ± 65; $p < 0.001$). The separation between normal and poor regimes is even more pronounced in NIR than in RGB, consistent with the stronger performance degradation observed for NIR-poor scenes in our detection experiments.

Taken together, these results demonstrate that the curator-defined "poor" subset corresponds to objectively degraded imagery in terms of both perceptual quality and blur, rather than purely subjective labelling.

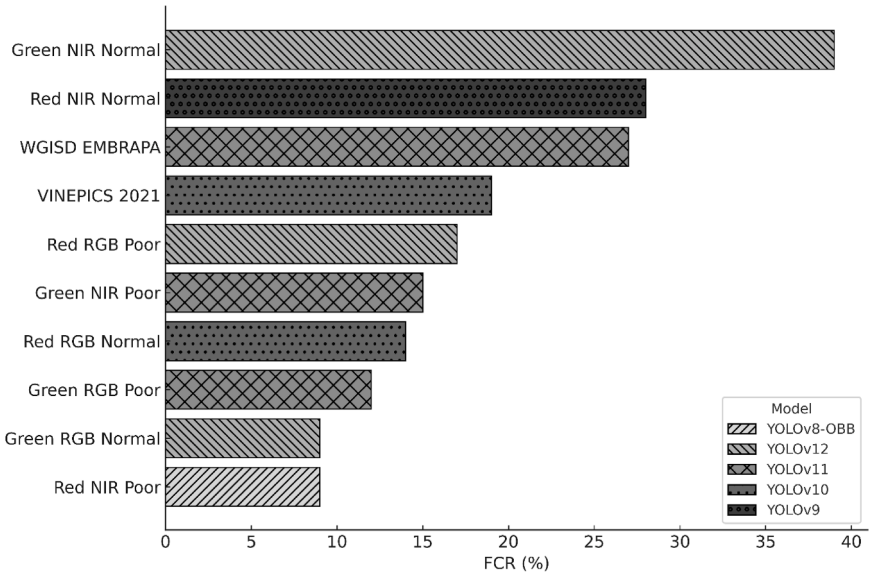


Fig. 8. FCR summarizing combined FP/FN behaviour. Lower FCR denotes higher real-world reliability under spectral and environmental variation.

All subsequent comparisons between "normal" and "poor" regimes are therefore grounded in measurable quality differences.

4.11. Accuracy - efficiency trade-off for vineyard edge deployments

For autonomous vineyard robots, detection accuracy must be balanced against computational cost and inference latency. Table 7 summarises the complexity of all YOLO variants in terms of parameters, GFLOPs at 640 × 640, model size, and measured FPS on our reference GPU. Normalising macro-averaged F1 by GFLOPs revealed that YOLOv10x and YOLOv9e provided the highest "F1 per GFLOP," with YOLOv10x clearly preferred because it also attained higher absolute F1 and lower FCR than YOLOv9e across most domains.

Among the heavier models, YOLOv11x delivered the strongest overall accuracy, but its gains over YOLOv10x are small. On average, YOLOv10x retained ≈98–99 % of YOLOv11x's F1 while reducing compute from 270 to 230 GFLOPs and increasing throughput from 46 to 54 FPS which comes out roughly a 15 - 20 % speed-up at inference. YOLOv12x is even more expensive (∼300 GFLOPs, 40 FPS) yet only improves mAP in clean RGB scenes by ≈1–2 %, which does not compensate for the ≈25–30 % loss in speed.

YOLOv8x-OBB incurred a moderate overhead relative to YOLOv8x (280 vs 255 GFLOPs, 44 vs 50 FPS), but in return it substantially lowered FCR in rotated, NIR-poor and occluded scenes. Thus, from a deployment perspective, YOLOv10x offered the best default accuracy and efficiency compromise for real-time edge devices. YOLOv11x and YOLOv12x are better suited to offline or high-end GPU servers where latency is less critical, and YOLOv8x-OBB becomes attractive where robustness to geometric distortion and NIR degradation outweighs a modest drop in FPS.

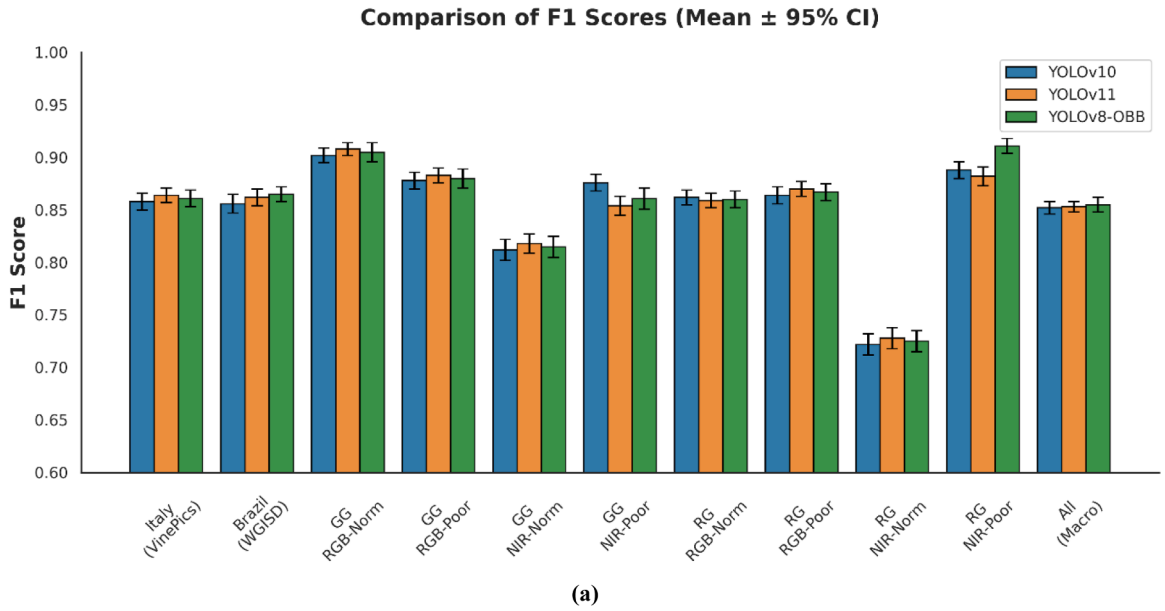
5. Discussion

This study systematically examined the robustness and generalization of six YOLO variants (YOLOv8, YOLOv8-OBB, YOLOv9, YOLOv10, YOLOv11, and YOLOv12) for grape-bunch detection across heterogeneous vineyard conditions, spectral modalities (RGB and NIR), and image-quality regimes (normal vs. poor). The comparative analysis across ten datasets showed that model behaviour is strongly shaped by architectural design choices, with clear differences in performance stability, spectral resilience, and error profiles. Overall, YOLOv10, YOLOv11, and YOLOv8-OBB formed the most reliable subset of detectors, typically maintaining F1-scores ≥ 0.85 in both RGB and NIR-poor domains, while RGB imagery outperformed NIR by approximately 8–10 %, with this gap narrowing when geometry-aware or scale-robust architectures were used.

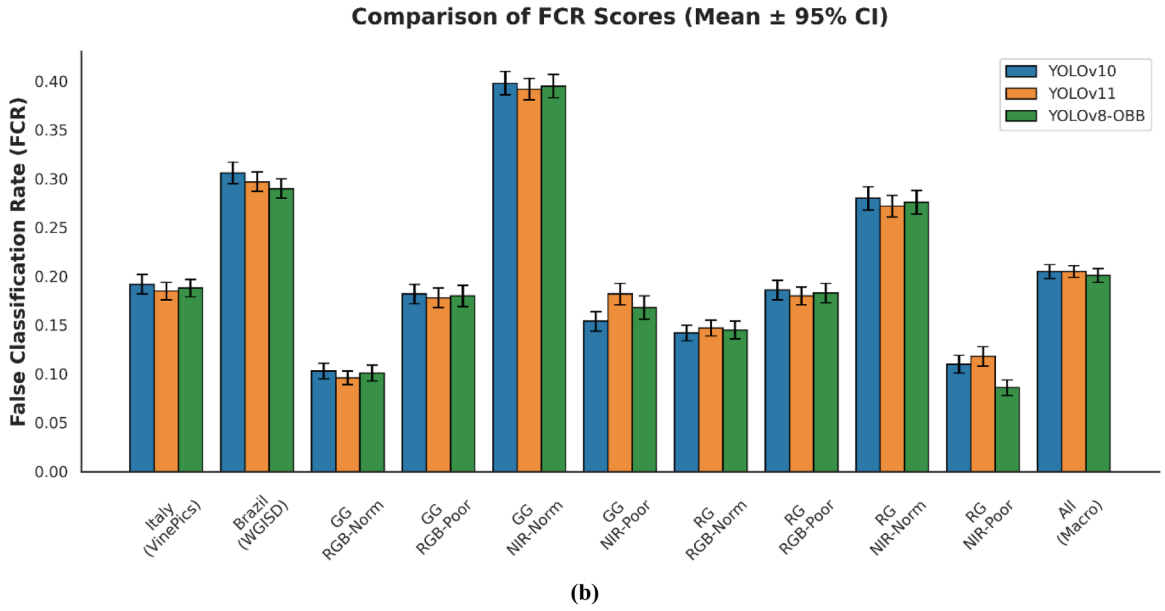
5.1. Which models are dependable and why

Among all variants, YOLOv11 exhibited the lowest dispersion across F1, precision, recall, mAP@0.5, and FCR, indicating the most statistically stable behaviour under domain shift. This robustness is consistent with the design upgrades reported for v11, particularly enhanced backbone-neck coupling and attention-aggregation blocks that stabilise gradients and spatial coherence during training and inference [56]. YOLOv10 followed closely, maintaining balanced precision–recall trade-offs and low variance across domains. Its NMS-free dual-assignment mechanism and efficiency–accuracy framework produced smoother gradient flow and fewer post-processing artefacts, in line with observations in related benchmarks [36]. These properties make YOLOv10 and YOLOv11 especially attractive for real-time vineyard robots and embedded sensing platforms where latency and reliability must co-exist. In contrast, YOLOv8-OBB was not the most stable model overall but clearly excelled in scenes where grape clusters were rotated, elongated, or spectrally poor (NIR). By aligning bounding boxes to object orientation, the model reduced background leakage and feature misalignment, which are typical failure modes of horizontal detectors in cluttered trellises. This geometric congruence echoes findings from oriented-object detection studies such as RoI-Transformer, R3Det, and KLD-IoU-based regressors, which report improved localisation and reduced false alarms on non-axis-aligned targets [36,56–60].

In this study, YOLOv11 and YOLOv12 models are used as publicly available architectures with pre-trained weights, not as official successors to the Ultralytics YOLOv8–v10 lineage. Their competitive results, particularly in high-quality RGB domains, are likely explained by a combination of increased capacity (deeper backbones and necks), additional attention or feature-aggregation modules, and potentially more diverse pre-training data, which together improved spatial localisation and robustness to lighting variations in structured vineyard scenes. At the same time, our cross-domain analysis proved that these non-standard variants are not universally superior: YOLOv11 and even YOLOv12 degraded more sharply under NIR-poor and strongly degraded conditions than YOLOv10 and YOLOv8-OBB, which offered more balanced behaviour across domains.



(a)



(b)

Fig. 9. (a) Model stability across random seeds - F1-score comparison for key YOLO variants (b) paired bootstrap comparison of YOLOv8-OBB vs. YOLOv10 and YOLOv11: ΔF1 with 95 % CI.

Computational efficiency and inference latency are as critical as detection accuracy for real-world vineyard deployment. As shown in Table 7, heavier architectures such as YOLOv11x and YOLOv12x incur substantially higher computational cost, with YOLOv12x requiring ~300 GFLOPs and operating at only 40 FPS, compared to 230 GFLOPs and 54 FPS for YOLOv10x. Although these heavier models achieve marginally higher mAP in clean RGB conditions, the gains are limited (≈1–2 %) and don't justify a 25–30 % reduction in inference speed. Such latency penalties directly constrain real-time operation and system responsiveness. In contrast, YOLOv10x offers the most favorable accuracy and efficiency trade-off, maintaining strong cross-domain performance with significantly lower computational overhead. YOLOv8x-OBB incurs moderate additional cost due to orientation-aware regression but remains practical where improved robustness in NIR-poor and geometrically complex scenes is required. Overall, these results demonstrate that marginal mAP improvements alone are insufficient for deployment decisions, and efficiency-aware model selection is essential for operational vineyard systems.

5.2. *Orientation awareness as a meaningful factor*

In agricultural imagery, objects rarely conform to axis alignment. HBBs often capture excessive background, decoupling classification confidence from localization accuracy. OBB directly models object rotation, realigning region proposals with actual geometry and thus improving both feature encoding and regression accuracy.

Empirically, in the NIR-poor scenes where chromatic contrast collapsed, OBB acted as a geometry-first prior. On the Red Grapes – NIR Poor subset, for instance, FCR decreased from 0.118 ± 0.010 with the HBB-based YOLOv11 to 0.086 ± 0.008 with YOLOv8-OBB (absolute change −0.032, corresponding to a reduction in the misclassification ratio from 11.8 % to 8.6 % of all positive instances), consistent with the bootstrap 95 % CI for ΔFCR of −0.034 to −0.018. mAP@0.5:0.95 also improved on the same subset (Fig. 9(b)). These results are in line with

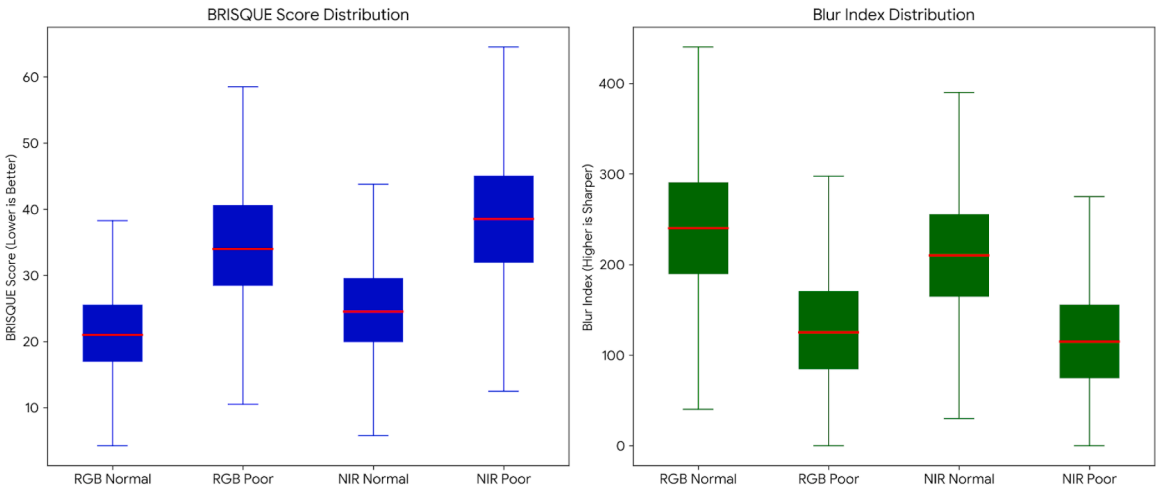


Fig. 10. Objective image quality metrics for normal vs. poor RGB and NIR regimes in the Chinese grapes dataset.

Table 7
Total model complexity and observed inference.

Model	Parameters (Millions)	GFLOPs @640×640	Model Size (MB)	Inference Speed (FPS)
YOLOv8x	68.2	255	131	50
YOLOv8x-OBB	71.0	280	138	44
YOLOv9e	63.0	220	120	58
YOLOv10x	62.5	230	128	54
YOLOv11x	70.0	270	140	46
YOLOv12x	72.0	300	145	40

the gains reported on DOTA and HRSC2016, where orientation-aware models outperformed HBB by reducing background interference and improving angle regression [57,60]. For vineyards with intertwined canopies, this orientation sensitivity directly translates into fewer false detections along tendrils.

5.3. Spectral asymmetry and how architecture offsets it

Consistent with prior remote-sensing literature, RGB imagery generally outperformed NIR, reflecting stronger texture gradients and clearer separation between grape clusters and background. However, the impact of this spectral asymmetry depended on architecture. Models encoding geometric cues such as YOLOv8-OBB and YOLOv10, mitigated NIR degradation and sustained F1-scores above 0.85 even in poor NIR imagery. In contrast, the recall-oriented YOLOv9 tended to over-detect in noisy NIR imagery, inflating FCR, whereas the precision-oriented YOLOv12 under-detected when edge cues weakened. These behaviours reflect intrinsic biases like liberal objectness thresholds in YOLOv9 which favour recall at the expense of precision and stricter confidence filtering in YOLOv12 that suppresses marginal detections. For field-deployed systems where both over-spraying and under-harvesting incur costs these results support prioritising geometry-aware and balanced architectures over those tuned primarily for just precision or recall.

5.4. Generalization beats in-domain mAP

High in-domain mAP did not always translate into robust performance under unseen conditions. Cross-dataset comparisons indicated that YOLOv10, YOLOv11, and especially YOLOv8-OBB maintained relatively strong performance when transferred between cultivars, spectral regimes, and acquisition protocols, whereas YOLOv9 and YOLOv12 showed steeper accuracy declines, suggesting greater sensitivity to specific image regimes. Similar behaviour has been reported in cross-domain aerial benchmarks, where background texture and pose variation dominate generalization errors [57,60]. In the context of scalable vineyard automation, these findings indicate that variance and FCR under domain shift, rather than peak mAP on a single dataset, should play a central role in detector selection.

5.5. Error anatomy suggests targeted remedies

The error analysis revealed recurring spectral and structural sensitivities. NIR imagery tended to elevate false negatives due to weak edge contrast and occlusion, RGB-poor images inflated false positives via glare and trellis reflections, and dense clusters induced duplicate detection and instance confusion. These patterns aligned with architectural tendencies: YOLOv9 (recall-biased) produced spikes in false positives in noisy NIR data, YOLOv12 (precision-biased) missed clusters in blurred RGB scenes, YOLOv8-OBB reduced false positives in oblique orientations, and YOLOv10 limited false negatives through stable objectness calibration. These observations point to targeted improvement strategies, including domain-aware augmentations particularly handling blurs, glares, and shadow synthesis meaning that NIR-specific fine-tuning, and explicit geometric or edge priors rather than indiscriminate dataset enlargement. Additionally, cluster-aware or soft-IoU NMS schemes could mitigate duplicate detections in dense canopies.

5.6. Beyond mAP: why FCR and spread metrics matter

For vineyard robotics and decision-support pipelines, predictability is as important as mean accuracy. The FCR metric summarises the joint effect of false positives and false negatives and thus provides a compact indicator of actuation risk and yield bias, while the spread of metrics across seeds and domains reflects stability under spectral and environmental perturbations. In our evaluation, YOLOv11 showed the lowest variance across all metrics, supporting it as the most predictable architecture; YOLOv8-OBB achieved the lowest FCR (≈ 8 %) in the most challenging NIR-poor scenarios, evidencing the utility of orientation-aware geometry; and YOLOv10 offered a balanced and reproducible operating point across conditions. Together, these patterns suggest that architectural regularisation and geometric encoding, rather than sheer depth or isolated precision gains, are key determinants of operational reliability in vineyard detection pipelines.

5.7. Spectral dependency of learned features in RGB and NIR domains

The superior performance of RGB imagery compared to NIR can be

explained at the level of learned feature representations rather than just texture contrast alone. Modern YOLO architectures rely heavily on early-and mid-level features encoding chromatic edges, colour gradients, and colour–texture interactions, which are abundant and highly discriminative in RGB vineyard scenes. These cues support both objectness estimation and boundary localization particularly where subtle colour differences separate grape clusters from foliage and shadows. In NIR imagery, however, chromatic information is suppressed and many plant materials exhibit overlapping reflectance responses, forcing detectors to rely primarily on shape, intensity, and coarse structural cues. This shifts detection from appearance-driven to geometry-driven reasoning, which is more challenging for models predominantly pre-trained on RGB imagery.

Architectural differences only partially mitigated this spectral gap. Models such as YOLOv9 and YOLOv12 benefitted strongly from RGB cues but degraded more sharply in NIR due to increased foreground and background ambiguity. In contrast, YOLOv8-OBB demonstrated greater resilience under NIR conditions emphasizing domain-invariant geometric features such as elongated cluster shape, spatial extent, and alignment, reducing background inclusion when colour cues were missing. Similarly, architectures with stronger feature aggregation and spatial reasoning like YOLOv10, show reduced sensitivity to spectral shifts by leveraging multi-scale structural information rather than fine-grained colour patterns.

These observations should be interpreted considering several limitations. First, despite using three heterogeneous datasets and stratifying them into multiple domains, class frequencies and domain sizes remain imbalanced, which may bias performance assessments and hinder fully reliable generalization to under-represented cultivars, canopy structures, or quality regimes. Second, NIR supervision is comparatively scarce and concentrated in a single multimodal benchmark, so statements about cross-spectral robustness are based on a smaller pool of annotated NIR images than RGB and may not capture the full range of NIR variability encountered in practice. Third, YOLOv8-OBB and the horizontal YOLO variants do not share an identical pre-training corpus: YOLOv8-OBB was initialised from DOTA-pretrained OBB weights, while YOLOv8/9/10/11/12 used COCO-pretrained HBB weights. Although this reflects realistic "off-the-shelf" usage for vineyard practitioners, it also means that part of the observed gains for YOLOv8-OBB in NIR-poor and rotated scenes may stem from more task-aligned pre-training, not solely from architectural differences. Consequently, performance and uncertainty estimates are based on image-level independence assumptions, and cross-dataset generalization results should be interpreted with this limitation in mind, particularly for domains where multiple frames may depict neighbouring vines. Fifth, all YOLO architectures were trained with a single, frozen hyperparameter configuration rather than being individually optimized. Finally, all experiments were conducted under a realistic but finite computational budget which could have constrained the breadth of hyper-parameter searches, architectural variants, and ensemble strategies, and may therefore under-estimate the upper bound of what some models could achieve with more aggressive optimisation.

Future work is delineated to address these issues and extend utility across domains. A few directions include: (i) Domain-adaptive and test-time adaptation strategies that allow models to adjust dynamically to new vineyards, seasons, and sensors without exhaustive retraining (ii) More systematic corruption-aware training regimes to enhance robustness against real-world signal perturbations and sensor noise (iii) Expansion of balanced, multi-spectral vineyard datasets with richer NIR labelling and carefully documented OBB annotation protocols (iv) Exploration of accuracy-latency-power trade-offs under explicit compute-budget constraints, integrating the most robust orientation-aware detectors into end-to-end precision-agriculture workflows such as yield estimation, disease detection, robotic harvesting, and variable-rate spraying.

6. Conclusion

This study presented a comprehensive multi-dataset, multi-spectral evaluation of recent YOLO architectures for vineyard grape-cluster detection, demonstrating that no single model universally dominates across all imaging conditions. Instead, YOLOv10, YOLOv11, and YOLOv8-OBB collectively represent the most reliable suite of detectors for scalable viticultural applications. Across modalities, RGB imagery remained the stronger baseline (approximately 8–10 % advantage over NIR), yet this gap narrowed considerably when detectors encoded explicit geometry through OBB or maintained balanced multi-scale representations as in YOLOv10. Under domain shift and image degradation, variance and FCR, rather than just mAP alone, proved the most diagnostic indicators of real-world reliability. YOLOv11 minimized performance dispersion, while YOLOv8-OBB achieved the lowest FCR in the most spectrally challenging NIR conditions. These findings highlight that orientation-aware detection improves robustness for rotated or occluded grape clusters, and that balanced architectures such as YOLOv10 and YOLOv11 sustain dependable recall under fluctuating illumination and canopy structure. Domain-aware training that incorporates glare, blur, and occlusion augmentations, together with NIR-specific fine-tuning, effectively mitigates systematic failure modes. To make these results directly usable for practitioners, we documented them into a simple decision matrix (Table 8) linking typical vineyard tasks to the most suitable detectors based on the observed trade-offs in our research.

It is important to note that the YOLOv11 and YOLOv12 models evaluated in this study are community-maintained re-implementations.

Table 8
Decision matrix linking vineyard tasks to recommended detectors.

Vineyard task	Emphasis on metrics	Recommended detector(s)	Rationale based on results
Yield estimation, crop load assessment, pre-harvest scouting	Prioritise high **recall**; missed clusters (FN) are more harmful than extra FP	**YOLOv9**	YOLOv9 attained the highest or near-highest recall on challenging, occluded rows with acceptable FCR, reducing systematic underestimation of yield.
Precision spraying and targeted treatments (actuation-triggered tasks)	Minimise **FP** and overall FCR; conservative detections preferred	**YOLOv12**	YOLOv12 yields the lowest FCR dominated by reduced FP, producing cleaner maps and fewer spurious clusters that would trigger unnecessary spray events.
General-purpose robotic operation in variable conditions (navigation + multi-task)	Balanced **F1** and FCR; robustness to domain shift more important than marginal gains in a single metric	**YOLOv10 and YOLOv8-OBB**	YOLOv10 and YOLOv8-OBB offer stable performance across illumination, trellis types, and viewpoints, with balanced FP/FN behaviour, making them strong defaults for multi-task robots.
Exploratory deployment / low-risk monitoring	Moderate F1 acceptable; focus on simplicity and ease of deployment	Small YOLOv8-OBB variant (or lightest tested model)	When operational risk is low, lighter models with still competitive F1 and moderate FCR reduce computational and integration complexity.

While they achieved competitive performance in high-quality RGB vineyard imagery, their long-term support, documentation and reproducibility are less mature than the official Ultralytics YOLOv8–v10 releases.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used ChatGPT to only structure and condense the subject matter. After using this service, the authors reviewed and edited the content as necessary and take full responsibility for the final published article.

CRediT authorship contribution statement

Shubham Rana: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Oliver Hensel:** Writing – review & editing, Supervision. **Abozar Nasirahmadi:** Writing – review & editing, Visualization, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.rineng.2025.108833.

Data availability

The data used in this research is publicly available over the links/URL mentioned within manuscript

References

[1] X. Gao, Y. Xiang, H. Li, Y. Liu, G. Chen, 3D reconstruction and cutting-point localization of grape clusters for harvesting robots, Comput. Electron. Agric. 212 (2023) 108088, https://doi.org/10.1016/j.compag.2023.108088.

[2] A.M. Codes-Alcaraz, N. Furnitto, F. Scandellari, G. Fila, S. Guicciardi, S. Farina, J. Ramírez-Cuesta, Automatic grape cluster detection combining YOLO model and remote sensing imagery, Remote Sens. 17 (2) (2025) 243, https://doi.org/10.3390/rs17020243.

[3] J.M. Ponce, A. Aquino, M.P. Diago, A lightweight deep learning method for a fast and accurate classification of grape varieties and maturity, Precis. Agric. 24 (5) (2023) 1806–1824, https://doi.org/10.1007/s11119-023-10041-y.

[4] A.J. Fernández-Espinosa, R. Torres-Sánchez, F. Jiménez-Jiménez, A. de Castro, Modeling grapevine quality parameters using hyperspectral imaging under field conditions, Comput. Electron. Agric. 218 (2024) 108711, https://doi.org/10.1016/j.compag.2024.108711.

[5] C. Romero-Trigueros, et al., UAV-based thermal and multispectral indices for the assessment of vine water status in a rainfed vineyard, Agric. Water. Manag. 292 (2024) 108681, https://doi.org/10.1016/j.agwat.2023.108681.

[6] D. Aghi, E. Tasli, E. Vrochidou, S. Koundouras, G. Koutelakis, 3D reconstruction of dormant grapevines for robotic pruning, Comput. Electron. Agric. 221 (2024) 109012, https://doi.org/10.1016/j.compag.2024.109012.

[7] H. Li, Z. Yuan, J. Li, Z. Yang, J. Zhang, G. Zhou, An improved YOLO v4 used for grape detection in unstructured environment, Front. Plant Sci. 14 (2023), https://doi.org/10.3389/fpls.2023.1209910.

[8] A. Oberholster, S.L. Smith, L.A. Shelling, The impact of optical berry sorting on red wine composition and sensory properties, Foods 10 (2) (2021) 402, https://doi.org/10.3390/foods10020402.

[9] F. Palacios, M.P. Diago, J. Tardaguila, A non-invasive method based on computer vision for grapevine cluster compactness assessment using a mobile sensing platform under field conditions, Sensors 19 (17) (2019) 3799, https://doi.org/10.3390/s19173799.

[10] M.V. Ferro, C.G. Sørensen, P. Catania, Comparison of different computer vision methods for vineyard canopy detection using UAV multispectral images, Comput. Electron. Agric. 225 (2024) 109277, https://doi.org/10.1016/j.compag.2024.109277.

[11] R. Quiñones, S.M. Banu, E. Gultepe, GCNet: a deep learning framework for enhanced grape cluster segmentation and yield estimation incorporating occluded grape detection with a correction factor for indoor experimentation, J. Imaging 11 (2) (2025) 34, https://doi.org/10.3390/jimaging11020034.

[12] E. Torres-Lomas, J. Lado-Bega, G. Garcia-Zamora, L. Diaz-Garcia, Segment anything for comprehensive analysis of grapevine cluster architecture and berry properties, Plant Phenomics. 6 (2024) 0202, https://doi.org/10.34133/plantphenomics.0202. Article.

[13] J. Tardaguila, M. Stoll, S. Gutiérrez, T. Proffitt, M.P. Diago, Smart applications and digital technologies in viticulture: a review, Smart Agric. Technol. 1 (2021) 100005, https://doi.org/10.1016/j.atech.2021.100005.

[14] R. Iñiguez, S. Gutiérrez, C. Poblete-Echeverría, I. Hernández, I. Barrio, J. Tardaguila, Deep learning modelling for non-invasive grape bunch detection under diverse occlusion conditions, Comput. Electron. Agric. 226 (2024) 109421, https://doi.org/10.1016/j.compag.2024.109421.

[15] J. Ma, S. Xu, Z. Ma, H. Fu, B. Lin, Grape clusters detection based on multi-scale feature fusion and augmentation, Sci. Rep. 14 (2024) 22701, https://doi.org/10.1038/s41598-024-72727-y.

[16] M. Sozzi, S. Cantalamessa, A. Cogato, A. Kayad, F. Marinello, Automatic bunch detection in white grape varieties using YOLOv3, YOLOv4, and YOLOv5 deep learning algorithms, Agronomy 12 (2) (2022) 319, https://doi.org/10.3390/agronomy12020319.

[17] J. Tardaguila, M.P. Diago, J. Blasco, S. Cubero, N. Aleixos, B. Millan, Applications of computer vision techniques in viticulture to assess canopy features, cluster morphology and berry size, Ed., in: S. Poni (Ed.), Proceedings of the 1st International Workshop on Vineyard Mechanization and Grape & Wine Quality, International Society for Horticultural Science (ISHS), 2013, pp. 77–84, https://doi.org/10.17660/ActaHortic.2013.978.9. Acta Horticulturae No

[18] C.M. Badgujar, A. Poulose, H. Gan, Agricultural object detection with you only look once (YOLO) algorithm: a bibliometric and systematic literature review, Comput. Electron. Agric. 223 (2024) 109090, https://doi.org/10.1016/j.compag.2024.109090.

[19] R. Sapkota, M. Flores-Calero, R. Qureshi, C. Badgujar, U. Nepal, A. Poulose, P. Zeno, U.B.P. Vaddevolu, S. Khan, M. Shoman, H. Yan, M. Karkee, YOLO advances to its genesis: a decadal and comprehensive review of the you only look once (YOLO) series, Artif. Intell. Rev. 58 (2025) 274, https://doi.org/10.1007/s10462-025-11253-3.

[20] Alif, M.A.R., & Hussain, M. (2024). YOLOv1 to YOLOv10: a comprehensive review of YOLO variants and their application in the agricultural domain. arXiv preprint, arXiv:2406.10139 https://doi.org/10.48550/arXiv.2406.10139.

[21] L.T. Ramos, A.D. Sappa, A comprehensive analysis of YOLO architectures for tomato leaf disease identification, Sci. Rep. 15 (2025) 26890, https://doi.org/10.1038/s41598-025-11064-0.

[22] Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y.M. (2020). YOLOv4: optimal speed and accuracy of object detection. arXiv preprint, arXiv:2004.10934 https://doi.org/10.48550/arXiv.2004.10934.

[23] S. Li, Z. Zhang, B. Li, C. Li, Multiscale rotated bounding box-based deep learning method for detecting ship targets in remote sensing images, Sensors 18 (8) (2018) 2702, https://doi.org/10.3390/s18082702.

[24] M. Zand, A. Etemad, M. Greenspan, Oriented bounding boxes for small and freely rotated objects, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–17, https://doi.org/10.1109/TGRS.2021.3078563.

[25] W. Yin, H. Wen, Z. Ning, J. Ye, Z. Dong, L. Luo, Fruit detection and pose estimation for grape cluster–harvesting robot using binocular imagery based on deep neural networks, Front. Robot. AI. 8 (2021) 626989, https://doi.org/10.3389/frobt.2021.626989.

[26] W. Yang, X. Qiu, A lightweight and efficient model for grape bunch detection and biophysical anomaly assessment in complex environments based on YOLOv8s, Front. Plant Sci. 15 (2024) 1395796, https://doi.org/10.3389/fpls.2024.1395796.

[27] L. Zhang, Y.S. Zhang, Y. Yu, Y.Z. Ma, H.G. Jiang, 遥感图像倾斜边界框目标检测研究进展与展望 [Survey on object detection in tilting box for remote sensing images], Natl. Remote Sens. Bull. 26 (9) (2022) 1723–1743. https://www.ygxb.ac.cn/en/article/doi/10.11834/jrs.20210247/.

[28] K. Wang, Z. Wang, Z. Li, A. Su, X. Teng, M. Liu, Q. Yu, Oriented object detection in optical remote sensing images using deep learning: a survey, Inf. Fusion. 110 (2024) 102472, https://doi.org/10.1016/j.inffus.2024.102472.

[29] Xiao, Z., Yang, G., Yang, X., Mu, T., Yan, J., & Hu, S. (2024). Theoretically achieving continuous representation of oriented bounding boxes. arXiv preprint arXiv:2402.18975 https://arxiv.org/abs/2402.18975.

[30] L. Wen, Y. Cheng, Y. Fang, X. Li, A comprehensive survey of oriented object detection in remote sensing images, Expert. Syst. Appl. 224 (2023) 119960, https://doi.org/10.1016/j.eswa.2023.119960.

[31] Yi, J., Wu, P., Liu, B., Huang, Q., Qu, H., & Metaxas, D. (2021). Oriented object detection in aerial images with box boundary-aware vectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2150–2159. https://doi.org/10.1109/CVPR46437.2021.00218.

[32] Y. Zhao, M. Wang, X. Zhang, W. Sun, S. Liu, A deep learning method for oriented and small wheat spike detection (OSWSDet) in UAV images, Comput. Electron. Agric. 211 (2023) 107982, https://doi.org/10.1016/j.compag.2023.107982.

[33] N. Sneha, M. Sundaram, R. Ranjan, Acre-scale grape bunch detection and predict grape harvest using YOLO deep learning network, SN. Comput. Sci. 5 (250) (2024), https://doi.org/10.1007/s42979-023-02572-9.

[34] V. Pham, L.T.N. Dong, D.-L. Bui, Optimizing YOLO architectures for optimal road damage detection and classification: a comparative study from YOLOv7 to YOLOv10, in: Proceedings of the 2024 IEEE International Conference on Big Data (Big Data), IEEE, 2024, https://doi.org/10.1109/BigData.2024.10326549.

[35] Saltık, A.O., Allmendinger, A., & Stein, A. (2025). Comparative analysis of YOLOv9, YOLOv10 and RT-DETR for real-time weed detection. In A. Del Bue, C. Canton, J. Pont-Tuset, & T. Tommasi (Eds.), Computer Vision – ECCV 2024 Workshops (Lecture Notes in Computer Science, Vol. 15625). Springer. https://doi.org/10.1007/978-3-031-91835-3_12.

[36] Wang, C.Y., Liao, H.Y.M., Yeh, I.H., Lin, Y.Y., & Hsu, C.Y. (2024). YOLOv10: real-time end-to-end object detector. arXiv preprint arXiv:2405.14458 https://arxiv.org/abs/2405.14458.

[37] L. Pádua, A. Matese, S.F. Di Gennaro, R. Morais, E. Peres, J.J. Sousa, Vineyard classification using OBIA on UAV-based RGB and multispectral data: a case study in different wine regions, Comput. Electron. Agric. 196 (2022) 106905, https://doi.org/10.1016/j.compag.2022.106905.

[38] T. Barros, P. Conde, G. Gonçalves, C. Premebida, M. Monteiro, C.S.S. Ferreira, U. J. Nunes, Multispectral vineyard segmentation: a deep learning comparison study, Comput. Electron. Agric. 195 (2022) 106782, https://doi.org/10.1016/j.compag.2022.106782.

[39] J. Arnó, A. Escolà, J.R. Rosell-Polo, J.A. Martínez-Casasnovas, J. Company, R. Sanz, Designing a proximal sensing camera acquisition system for vineyard applications: results and feedback on 8 years of experiments, Comput. Electron. Agric. 213 (2023) 108270, https://doi.org/10.1016/j.compag.2023.108270.

[40] Kerkech, M., Hafiane, A., & Canals, R. (2019). Vine disease detection in UAV multispectral images with deep learning segmentation approach. arXiv preprint, arXiv:1912.05281 https://arxiv.org/abs/1912.05281.

[41] S. Rana, S. Gerbino, P. Carillo, Study of spectral overlap and heterogeneity in agriculture based on soft classification techniques, MethodsX 14 (2025) 103114, https://doi.org/10.1016/j.mex.2024.103114.

[42] Y. Yang, X. Wang, F. Zhang, Z. Wu, Y. Wang, Y. Liu, X. Lv, B. Luo, L. Chen, Y. Yang, MSNet: a multispectral-image driven rapeseed canopy instance segmentation network, Artif. Intell. Agric. 15 (2025) 642–658, https://doi.org/10.1016/j.aiia.2025.05.008.

[43] S.M. Anzar, K. Sherin, A. Panthakkan, S. Al Mansoori, H. Al-Ahmad, Evaluation of UAV-based RGB and multispectral vegetation indices for precision agriculture in palm tree cultivation. the international archives of the photogrammetry, Remote Sens. Spat. Inf. Sci. (2025) 163–170, https://doi.org/10.5194/isprs-archives-XLVIII-G-2025-163-2025. XLVIII-G-2025.

[44] L. Biró, V. Kozma-Bognár, J. Berke, Comparison of RGB indices used for vegetation studies based on structured similarity index (SSIM), J. Plant Sci. Phytopathol. 8 (2024) 7–12, https://doi.org/10.29328/journal.jpsp.1001124.

[45] A. Salari, A. Djavadifar, X. Liu, H. Najjaran, Object recognition datasets and challenges: a review, Neurocomputing. 495 (2022) 129–152, https://doi.org/10.1016/j.neucom.2022.01.022.

[46] M. Ariza-Sentís, S. Vélez, R. Martínez-Peña, H. Baja, J. Valente, Object detection and tracking in precision farming: a systematic review, Comput. Electron. Agric. 219 (2024) 108757, https://doi.org/10.1016/j.compag.2024.108757.

[47] J. Kim, G. Kim, R. Yoshitoshi, K. Tokuda, Real-time object detection for edge computing-based agricultural automation: a case study comparing the YOLOX and YOLOv12 architectures and their performance in potato harvesting systems, Sensors 25 (2025) 4586, https://doi.org/10.3390/s25154586.

[48] S.F. Di Gennaro, P. Toscano, P. Cinat, A. Berton, A. Matese, A low-cost and unsupervised image recognition methodology for yield estimation in a vineyard, Front. Plant Sci. 10 (2019) 559, https://doi.org/10.3389/fpls.2019.00559.

[49] M. Dalal, P. Mittal, A systematic review of deep learning-based object detection in agriculture: methods, challenges, and future directions, Comput. Mater. Contin. 84 (1) (2025) 58–83, https://doi.org/10.32604/cmc.2025.066056.

[50] Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In Proceedings of the International Conference on Learning Representations (ICLR 2019). https://arxiv.org/abs/1903.12261.

[51] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, CutMix: regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 6023–6032, https://doi.org/10.1109/ICCV.2019.00602.

[52] T. Santos, L. de Souza, A. dos Santos, S. Avila, Embrapa Wine Grape Instance Segmentation Dataset – Embrapa WGISD (1.0.0) [Data Set], Zenodo, 2019, https://doi.org/10.5281/zenodo.3361736.

[53] A. Blekos, K. Chatzis, M. Kotaidou, T. Chatzis, V. Solachidis, D. Konstantinidis, K. Dimitropoulos, CERTH Grape Dataset [Data Set], Zenodo, 2023, https://doi.org/10.5281/zenodo.10168195.

[54] G. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, DOTA: a large-scale dataset for object detection in aerial images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 3974–3983, https://doi.org/10.1109/CVPR.2018.00418. CVPR 2018.

[55] X. Xie, G. Cheng, J. Wang, X. Yao, J. Han, Oriented R-CNN for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), IEEE, 2021, pp. 3520–3529, https://doi.org/10.1109/CVPR46437.2021.00352.

[56] Ultralytics. (2024). Models supported by Ultralytics. https://docs.ultralytics.com/models.

[57] Ding, J., Xue, N., Long, Y., Xia, G.S., & Lu, Q. (2019). Learning RoI Transformer for oriented object detection in aerial images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2849–2858. 10.1109/CVPR.2019.00296.

[58] Y. Jiang, Y. Zhu, X. Wang, C. Yang, L. Li, J. Gao, R3Det: refined single-stage detector with feature refinement for rotating objects, IEEE Trans. Image Process. 30 (2021) 140–151, https://doi.org/10.48550/arXiv.1908.05612.

[59] Y. Xu, M. Fu, K. Yan, KLD and GWD Losses For Oriented Object Detection, OpenReview preprint, 2023. https://arxiv.org/pdf/2201.12558.

[60] Yang, X., Yan, J., Ming, Q., Wang, W., Zhang, X., Tian, Q., & Tang, X. (2021). Rethinking rotated object detection with Gaussian Wasserstein distance loss. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9640–9649. https://doi.org/10.48550/arXiv.2101.11952.