International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 5, Number: 1-1, Pages: 1 - 10, Year: 2026

Multi-Defect Classification in Oranges Using YOLOv8, YOLO11, and YOLOv8m: A Comparative Study

Muhammad Daniyal Baig[1], Shafqat Hussain[2], Sidrah Aslam[3], Sabir Abbas[4], Amna Kosar[5], Imran Khalid[6], Hafiz Burhan Ul Haq[7,*], Malik Hammad Hussain[8]

[1,8]NASTP NIIT, Lahore, Pakistan
[2,5]Department of Computer Science, Faculty of Computer Sciences, Lahore Garrison University, Lahore, 54000, Pakistan
[3,4,7]Department of Artificial Intelligence, Faculty of Computer Sciences, Lahore Garrison University, Lahore, 54000, Pakistan
[5]Department of Criminology and Forensic Science, Faculty of Social Sciences, Lahore Garrison University, Lahore, 54000, Pakistan
Corresponding Author: Hafiz Burhan Ul Haq (Email: burhanhashmi64@lgu.edu.pk)

Received: 12/05/2026, Revised: 22/07/2026, Accepted: 06/08/2026

Abstract—Orange farming is among the most economically significant agricultural activities in the world, with the crop feeding global markets and supporting millions of smallholder livelihoods. Yet despite that importance, quality grading at the post-harvest stage still depends heavily on manual inspection, a slow and inconsistent process that struggles to keep pace with the demands of modern supply chains. Surface diseases such as Blackspot, Canker, and Greening are among the leading causes of post-harvest losses, and catching them early is critical to protecting fruit quality and commercial value. This paper presents an automated classification framework for orange fruit defects that evaluates three members of the You Only Look Once (YOLO) model family: YOLOv8n, YOLO11n, and an enhanced YOLOv8m trained with an extended preprocessing and augmentation pipeline. The system classifies orange images into four categories—Blackspot, Canker, Fresh, and Greening—enabling defect-specific grading decisions rather than simple pass/fail quality labelling. All models were trained and tested on a publicly available orange disease dataset under identical experimental conditions, providing a controlled multi-model benchmark. Evaluation metrics include Top-1 Accuracy, Precision, Recall, and F1-Score per class. The enhanced YOLOv8m achieved the best overall performance with 97.98% accuracy and a macro F1-Score of 0.98, followed by YOLO11n at 96.97% and YOLOv8n at 93.94%. These results show that targeted augmentation and regularization strategies can produce meaningful gains in classification performance even on small, class-imbalanced agricultural datasets.

Index Terms—Orange Disease Classification, Defect Detection, YOLOv8, YOLO11, Deep Learning, Precision Agriculture, Transfer Learning

I. INTRODUCTION

Oranges are one of the most widely consumed fruits globally, valued for their nutritional profile, flavour, and versatility across food and beverage industries. Production runs into hundreds of millions of metric tons annually, spanning tropical and subtropical regions across Asia, the Americas, the Mediterranean, and Africa. Beyond large commercial growers, citrus cultivation sustains millions of smallholder farming households for whom consistent fruit quality is directly tied to income and livelihood security. Post-harvest quality management, however, presents persistent challenges. Oranges are susceptible to several surface diseases and defects that reduce market value, shorten shelf life, and can render batches unsuitable for export. The most economically damaging of these are Blackspot, a fungal infection that produces dark lesions on the fruit surface; Canker, a bacterial disease causing raised, corky growths with characteristic yellow halos; and Greening, a systemic bacterial condition transmitted by insects that leads to uneven coloration and poor flavor development. When affected fruit enters the supply chain undetected, it accelerates spoilage of surrounding produce, triggers post-export rejections, and generates losses across the entire value chain. Current practice at most packing houses relies on trained human inspectors who examine fruit visually as it moves along grading lines. While familiar and relatively inexpensive in terms of equipment, manual inspection has well-documented limitations. Inspectors are subject to fatigue, perceptual variability, and inconsistent grading standards. Early-stage defects that have not yet developed full visual expression are especially easy to miss, meaning that affected fruit often progresses through the chain until the damage is undeniable, by which point substantial losses have already occurred.

Computer vision and deep learning have, over the past decade, offered compelling alternatives to manual inspection across agricultural quality assessment. Convolutional Neural Networks (CNNs) have demonstrated strong performance on image-based crop disease classification tasks, and the YOLO family of real-time models has emerged as a particularly practical option for deployment environments that demand both speed and accuracy. YOLO models perform classification in a single forward pass through the network, enabling high throughput without sacrificing classification fidelity. This study proposes an automated defect classification framework for orange fruit using three YOLO variants: YOLOv8n as a lightweight baseline, YOLO11n as the latest stable release with native classification support, and an enhanced YOLOv8m trained with an extended augmentation and regularization pipeline specifically designed to address the class imbalance and visual similarity challenges inherent in orange disease datasets. The four-class task—Blackspot, Canker, Fresh, and Greening—supports defect-type-specific grading decisions rather than broad binary quality labelling. Beyond the leaf-level and architecture-focused work discussed later in this paper, several recent studies have specifically targeted deep-learning-based citrus disease classification. Butt et al. [1] combined transfer learning with a hybrid meta-heuristic feature-selection strategy to classify citrus leaf diseases, while Zhang et al. [2] proposed a hybrid attention network for the same task. Goyal and Lakhwani [3] benchmarked four CNN architectures across nine citrus disease classes spanning both leaves and fruit, and Mai et al. [4] provide a broader review of machine-learning workflows for citrus quality and disease detection. Peng et al. [5] further survey machine-vision applications across the citrus production pipeline, including disease detection, harvesting, and grading. Collectively, this body of work underscores the growing role of deep learning in citrus disease management, motivating the present study's focus on fruit-level defect detection using YOLO-based object detectors.

While the CNN- and transfer-learning-based studies reviewed above (references [18]-[21]) report high classification accuracy on the same or closely related Kaggle-sourced orange disease dataset, they operate exclusively at the level of whole-image classification and do not benchmark against modern single-stage object-detection architectures. The novelty of the present study lies in (i) reframing this four-class orange defect problem as a YOLO-based classification task, which naturally extends to spatial defect localization in future work; (ii) providing a controlled, head-to-head comparison of three YOLO-family variants-YOLOv8n, YOLO11n, and an enhanced YOLOv8m-trained and evaluated under identical experimental conditions; and (iii) introducing an extended augmentation and regularization pipeline targeted specifically at the Blackspot-Canker visual-similarity challenge, which prior CNN-based studies on this dataset do not address.

II. LITERATURE REVIEW

A. Background: Citrus Disease and Defect Detection

Oranges are among the most economically important fruits worldwide, yet citrus orchards remain persistently vulnerable to fungal and bacterial pathogens that produce visible defects on the fruit rind, including orange canker, black spot, melanose, and many more. These conditions degrade fruit appearance, reduce marketable yield, and impose substantial economic losses on farmers, particularly where manual sorting and inspection remain the dominant grading practice. Manual inspection is labor-intensive, subjective, and difficult to scale to the throughput demanded by modern packing houses, which has motivated several decades of research into computer-vision-based alternatives. Early approaches relied on handcrafted color, texture, and shape descriptors coupled with classical classifiers such as support vector machines and k-means clustering; while computationally inexpensive, these pipelines proved brittle under varying illumination, fruit orientation, and disease severity, and required extensive manual feature engineering.

B. YOLO Architectures for Object Detection

The YOLO family has progressed steadily in object detection since its inception, with each successive version targeting a different balance of accuracy, latency, and model size. Terven et al. [6] give a comprehensive architectural review of this evolution from YOLOv1 through YOLOv8, documenting the field's gradual shift from anchor-based two-stage detection pipelines toward anchor-free, single-stage designs with decoupled classification and regression heads. Ranjan et al. [7] provide a complementary decadal review that similarly traces how successive versions have refined the backbone, neck, and head components to improve multi-scale feature fusion while reducing inference cost. YOLOv8, released by Ultralytics in 2023, replaced the earlier CSP layer with the C2f (cross-stage partial bottleneck with two convolutions) module, adopted an anchor-free detection head, and unified detection, segmentation, classification, and pose-estimation tasks within a single framework offered in five scaled variants: nano, small, medium, large, and extra-large. Terven et al. [6] further discuss YOLOv8m, the medium-scale variant evaluated in this study, which sits at the midpoint of this scaling spectrum, providing greater representational capacity than the nano and small configurations while remaining considerably lighter than the large and extra-large variants, making it an attractive candidate for orchard and packing-line deployable inspection systems. Sapkota et al. [15] discuss YOLO11, the more recent architectural generation evaluated alongside YOLOv8 in this study.

C. YOLOv8-Based Detection of Citrus and Orange Diseases

YOLOv8 has been widely adapted for citrus disease detection, although the bulk of this work targets leaf-level rather than fruit-level symptoms. Zheng et al. [8] proposed an enhanced YOLOv8n variant incorporating a multi-scale attention fusion module in the neck to address the visual similarity between citrus leaf disease classes, reporting a 1.2 percentage point accuracy gain alongside a 7% reduction in parameter count relative to the baseline. Zhu et al. [9] combined CSPPC multi-dimensional and spatial-convolution modules into a YOLOv8-CMS architecture for joint citrus leaf disease classification and severity grading, benchmarking against YOLOv5, YOLOv3, and multiple YOLOv8 configurations. Targeting the economically devastating Huanglongbing disease specifically, Xie et al. [10] introduced YOLO-EAF, which augments YOLOv8n with an efficient multi-scale attention module and an asymptotic feature fusion strategy in the neck, improving mAP@0.5 by 1.9 percentage points over the unmodified backbone on small, visually subtle lesions. Other lightweight reformulations of YOLOv8 for citrus disease detection include a dilated re-parameterization feature enhancement module paired with a shared-parameter detection head proposed by Guo et al. [11], which reduced parameter count by roughly 48% and computational load by close to 59% relative to baseline YOLOv8 while achieving 97.3% mAP, and a slim-neck design using GSConv and VoV-GSCSP blocks together with an MLCA attention mechanism for joint pest and disease detection discussed by Wang et al. [12].

Beyond leaf-level diagnosis, a smaller body of work has applied YOLOv8 directly to fruit-level quality assessment. Than et al. [13] compared YOLOv8 against the lightweight NanoDet detector for real-time orange fruit detection and quality evaluation on mobile devices, demonstrating that single-stage detectors can support on-device fruit screening applications. Joshi [14] evaluated a YOLOv8-S configuration for orange species and disease detection, covering citrus canker, black spot, sooty mould, blue-green mould, and citrus greening, under resource-constrained edge-computing conditions, reporting an mAP@0.5 of 0.949 alongside low per-image inference latency, underscoring YOLOv8's suitability for field-deployable orange screening pipelines. Miao et al. [26] explore the broader applicability of YOLOv8 to plant disease detection beyond citrus, further evidenced by enhancements such as SerpensGate-YOLOv8 for general plant disease detection and YOLOv8-G for dragon-fruit stem disease detection explored by Gómez-Flores et al. [27], both of which modify the C2f module.

D. YOLO11 and Comparative Evaluations of Recent YOLO Variants in Agricultural Imaging

Because YOLO11 was released only recently, the agricultural literature addressing it remains largely comparative, typically benchmarking it directly against YOLOv8 rather than studying it in isolation. Sapkota and Karkee [15] compared YOLO11 and YOLOv8 for instance segmentation of occluded and non-occluded immature green fruit in commercial apple orchards, reporting that YOLO11's refined architectural and training optimizations improve processing efficiency relative to YOLOv8, although the magnitude of improvement varied with the degree of fruit occlusion. In a fruit-counting context, [16] compared the nano and extra-large configurations of YOLOv8 and YOLO11 for tart-cherry detection on a mechanical harvester, finding that YOLO11x achieved a marginally higher mAP@0.5 (0.92) than the corresponding YOLOv8 configuration, while the nano variants performed similarly, with YOLOv8n offering faster inference. The most extensive comparative benchmark to date, [17], evaluated YOLOv8, YOLOv9, YOLOv10, YOLO11, and YOLOv12 for fruitlet detection and counting across multiple apple cultivars and imaging sensors, finding that no single architecture dominates uniformly across precision, recall, mAP, and inference-efficiency metrics; YOLO11 variants achieved competitive mAP@0.5 scores and the fastest inference among the tested models.

E. CNN and Transfer-Learning-Based Classification of Orange Diseases

A parallel and substantially larger body of literature has addressed orange disease identification as a whole-image classification problem using CNNs and transfer learning, frequently on the same or closely related Kaggle-sourced orange disease imagery used in the present study. Reference [18] proposed a CNN model combined with color-segmentation preprocessing and Gaussian/median-filter-based denoising for classifying 1,790 Kaggle-sourced orange images into greening, canker, black spot, and fresh categories. Reference [19] fine-tuned a MobileNetV2 backbone with additional dropout and average-pooling layers to classify the same four disease categories from roughly 1,090 images, reporting 95% validation accuracy. Reference [20] extracted CNN-based features from both lemon and orange disease datasets and passed them to conventional machine-learning classifiers, again using the four-class dataset originally released by Silva [29]. Reference [21] combined a CNN feature extractor with a gradient-boosting classifier in a hybrid multi-optimizer framework for a citrus fruit dataset of roughly 3,000 images. More recently, [22] proposed a dual-branch attention-augmented network together with explainable-AI visualization for disease classification across several fruit types, including an orange subset, reporting strong classification accuracy alongside saliency-based interpretability. While these CNN- and transfer-learning-based studies achieve high classification accuracy on cropped, single-fruit images, none of them perform spatial localization of the defect within the frame, a capability that single-stage object detectors such as YOLOv8 and YOLO11 provide.

F. Systematic Reviews and Surveys on Citrus and Fruit Quality Assessment

Several systematic reviews contextualize the methodological landscape surrounding this study. Dhiman et al. [23] systematically reviewed image-acquisition protocols, preprocessing pipelines, and classification approaches reported across the citrus fruit disease detection literature, concluding that dataset diversity, illumination control, and the choice between classification- and detection-based problem formulations remain key sources of variability in reported accuracy across studies. A broader systematic review of deep-learning-based object detection methods for crop grading [24] traces the field's progression from earlier YOLO versions toward designs emphasizing anchor-free detection, decoupled heads, and dynamic label assignment, noting that such design choices improve sensitivity to the fine-grained surface defects relevant to fruit grading tasks. A further review specifically assessing YOLO-based plant disease detection [25] highlights the consistent gains reported when YOLO detectors are applied to fruit and leaf disease identification relative to earlier two-stage and purely classification-based pipelines.

Table I summarizes the studies reviewed above. Taken together, the literature reveals three consistent patterns. First, YOLOv8 has been extensively and successfully adapted for citrus disease detection, but the overwhelming majority of this work targets leaf-level disease symptoms, such as Huanglongbing and canker-induced leaf-spotting, rather than fruit-surface defects, and almost always proposes a modified or "improved" YOLOv8 variant rather than benchmarking the unmodified architecture against newer YOLO generations. Second, comparative evaluations of YOLO11 against YOLOv8 are emerging rapidly in fruit-detection and fruit-counting contexts such as apples, cherries, and fruitlets, consistently showing that neither architecture dominates uniformly and that performance differences are dataset- and scale-dependent; however, no published comparison of this kind has, to the authors' knowledge, been conducted specifically for multi-class orange fruit defect detection. Third, the substantial body of CNN- and transfer-learning-based work on the same publicly available orange disease dataset used in this study has consistently treated the problem as whole-image classification, leaving the complementary task of simultaneous defect localization and multi-class classification comparatively underexplored. The present study addresses this gap by training and systematically comparing YOLOv8n, YOLOv8m, and YOLO11n on the same orange multi-defect dataset.

TABLE I. SUMMARY OF KEY STUDIES REVIEWED

Ref.	Model(s)	Task / Crop
[6]	Architectural review (YOLOv1–v8, YOLO-NAS)	Survey of object detection architectures
[8]	Improved YOLOv8n + multi-scale attention fusion	Citrus leaf disease detection
[9]	YOLOv8-CMS (CSPPC, MultiDimen, SpatialConv)	Citrus leaf disease classification & grading
[10]	YOLO-EAF (YOLOv8n + EMA + ASFF)	Citrus Huanglongbing (greening) detection

III. DATASET

A. Data Source

The dataset used in this study is a publicly available orange disease image collection sourced from Kaggle, containing images of orange fruit across four classes: Blackspot, Canker, Fresh, and Greening. Images are organized in class-specific subdirectories following the standard ImageFolder convention, making the dataset directly compatible with both the Ultralytics YOLO training framework and PyTorch data loaders. The full dataset comprises 1,090 images: 991 in the training partition and 99 in the test partition. Within the training set, class distribution is as follows: Greening (347 images), Fresh (281 images), Blackspot (184 images), and Canker (179 images). This distribution reflects a moderate degree of class imbalance, with the Greening and Fresh classes containing substantially more samples than Blackspot and Canker, a pattern that influenced the augmentation strategy applied to the enhanced YOLOv8m model. The test set contains 22 images each for Blackspot, Canker, and Greening, and 33 images for Fresh.

It should be noted that, relative to large-scale benchmark datasets used elsewhere in the deep-learning literature, this dataset is comparatively small, and the held-out test partition (99 images, 22-33 images per class) is likewise limited. Reported accuracy and F1-Score figures should therefore be interpreted with appropriate caution: performance estimates on such a small test set carry non-trivial sampling variance, and the reported gains between models, while consistent with the underlying architectural and training differences, should not be read as precise population-level estimates. Validation on larger and more diverse datasets, discussed further in Section VI, is necessary before these results can be generalized to real-world packing-house conditions.

Fig. 1. Representative sample images from the dataset: Greening, Fresh, Canker, and Blackspot.

B. Defect Classes

The four classification categories covered in this study are:

- **Fresh:** undamaged orange fruit with no visible surface abnormalities, representing normal produce suitable for market.
- **Blackspot:** dark, irregularly shaped lesions caused by fungal infection (Phyllosticta citricarpa), typically appearing as scattered spots on the peel surface.
- **Canker:** raised, corky lesions resulting from bacterial infection (Xanthomonas citri), usually surrounded by a yellow water-soaked halo.
- **Greening:** uneven surface coloration and poor fruit development caused by the systemic bacterial pathogen Candidatus Liberibacter, transmitted by the Asian citrus psyllid.

C. Validation Split

The original dataset does not include a predefined validation partition. To enable principled model selection during training, including early stopping, learning-rate scheduling, and monitoring of generalization performance, a held-out validation set was carved from the training data before any model training began. Fifteen percent of images from each class folder were randomly selected and moved to a dedicated validation directory, using a fixed random seed of 42 to ensure the split was deterministic and fully reproducible. This stratified sampling approach prevents class-imbalance artefacts in the validation set that could distort per-class validation metrics. The resulting partition is approximately 77.6% training, 13.4% validation, and 9% test across all four classes. The test set was withheld entirely from all stages of model development and used only for final performance evaluation.

D. Preprocessing and Augmentation

All images were resized to 224 × 224 pixels prior to training, consistent with the input requirements of pretrained YOLO classification backbones. Bilinear interpolation was used during resizing to preserve image detail and avoid aliasing artefacts. Pixel values were normalized using ImageNet population statistics (mean = [0.485, 0.456, 0.406]; std = [0.229, 0.224, 0.225] per RGB channel). This normalization step ensures that input distributions fall within the statistical regime for which the pretrained feature extractors were optimized, supporting effective knowledge transfer.

YOLOv8n and YOLO11n were trained with default Ultralytics augmentation settings, which include random horizontal flipping, mosaic augmentation, and HSV colour jitter. These transformations provide a baseline level of diversity without altering class-defining features. The enhanced YOLOv8m was trained with an extended augmentation configuration built to address the specific challenges of this dataset, particularly the class imbalance and the visual similarity between Blackspot and Canker. This extended configuration included:

- HSV hue, saturation, and brightness jitter (hsv_h = 0.015, hsv_s = 0.7, hsv_v = 0.4) to simulate natural variation in fruit surface colour and ambient illumination.

- Random horizontal flip (probability 0.5) and vertical flip (probability 0.1) to account for arbitrary fruit orientations in grading environments.
- Random rotation of up to ±15° and scale variation of 0.5 to simulate varying capture angles and imaging distances.
- Random erasing (probability 0.4) applied to rectangular patches to encourage distributed feature representations rather than reliance on localized texture cues.
- RandAugment automatic augmentation policy, sampling from a broad transformation space to further diversify the training distribution.
- Label smoothing ($\varepsilon = 0.1$) to reduce overconfidence on visually similar classes.
- Dropout regularization (rate = 0.3) within the classification head to reduce co-adaptation of learned features.

IV. METHODOLOGY

This study implements an end-to-end automated classification pipeline for orange fruit surface diseases. The pipeline progresses sequentially through dataset preparation, image preprocessing, augmentation strategy selection, model initialization, training configuration, and performance evaluation. All three models—YOLOv8n, YOLO11n, and the enhanced YOLOv8m—share identical conditions wherever their configurations overlap, with deliberate and clearly documented departures introduced only where the experimental design requires them. The classification problem is formulated as a four-class single-label image classification task. Each input image is assigned to one of four mutually exclusive categories: Blackspot, Canker, Fresh, or Greening. This framing is well-suited to post-harvest quality grading, where a sorting decision must be made for each fruit as a discrete unit. The system is not designed to localize defect regions within images; rather, it assigns a class-level quality label appropriate for downstream grading logic in packing-house environments.

A. Dataset Preparation

Dataset preparation followed the process described in Section III. After the validation split, the final data partition used for all modelling consisted of approximately 841 training images, 150 validation images, and 99 test images across the four classes. Images were stored in class-specific subdirectories compatible with the Ultralytics YOLO ImageFolder convention. The test set was isolated from all stages of training and hyperparameter selection.

B. Image Preprocessing

All images across training, validation, and test splits were passed through an identical preprocessing pipeline. Raw images were resized to 224 × 224 pixels using bilinear interpolation, then normalized using ImageNet mean and standard deviation values. Applying this preprocessing pipeline identically to all three models ensures that observed performance differences can be attributed to model architecture and augmentation strategy rather than to inconsistencies in input representation.

C. Data Augmentation Strategy

Augmentation was applied during training only and was not applied to validation or test images, preserving evaluation integrity. The augmentation strategy was deliberately differentiated across models to reflect the experimental design. YOLOv8n and YOLO11n were trained with the default Ultralytics augmentation configuration (Section III-D), while the enhanced YOLOv8m was trained with the extended pipeline designed specifically for this dataset's challenges. The rationale for applying the extended pipeline specifically to YOLOv8m is twofold. First, the medium-scale model has sufficient parameter capacity to exploit the additional training diversity introduced by aggressive augmentation without under-fitting. Second, the Blackspot and Canker classes share overlapping surface texture and color characteristics that standard augmentation alone does not adequately disentangle. The label smoothing, random erasing, and RandAugment components were selected with this inter-class confusion challenge explicitly in mind. Because this extended augmentation pipeline was combined with a larger backbone (YOLOv8m) rather than isolated as a controlled variable, the performance improvement attributed to YOLOv8m Enhanced in Section V reflects the joint effect of both factors; this design choice and its implications for interpreting the results are discussed further in Section V-C.

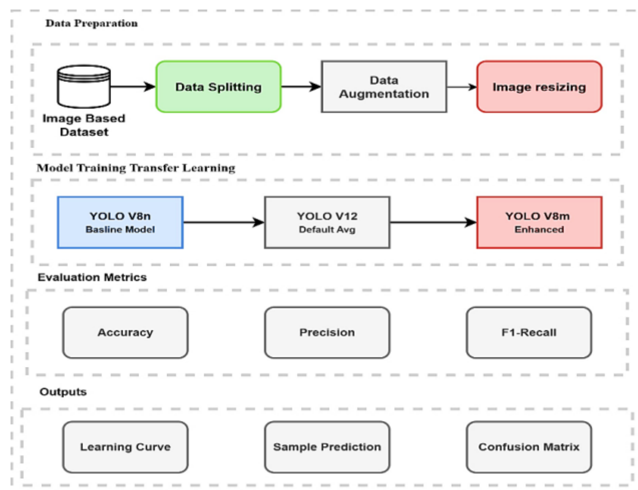

Fig. 2. Overview of the proposed methodology pipeline, from data acquisition through model training to comparative evaluation.

D. Classification Architecture

All three models in this study use the YOLO image classification architecture, which adapts the standard YOLO detection backbone for image-level prediction. The backbone processes each input image through a sequence of convolutional layers and C2f feature extraction blocks, culminating in a Spatial Pyramid Pooling Fast (SPPF) module that aggregates multi-scale contextual representations. The classification head replaces the standard bounding-box detection head with global average pooling, an optional dropout layer, and a fully connected linear layer that maps the pooled feature vector to a probability distribution over the four target classes via a softmax activation. This single-stage classification architecture suits the grading task for two reasons. First, since the objective is class assignment rather than defect localization, the added complexity and computational overhead of a detection head is unnecessary. Second, the end-to-end pipeline enables efficient inference suitable for high-throughput packing-house deployment.

YOLOv8n is the nano-scale variant of YOLOv8, offering the smallest parameter footprint in the family. It serves as the lightweight baseline in this study, establishing the classification accuracy achievable under constrained computational budgets with standard augmentation. YOLO11n is the nano-scale variant of YOLO11, the current stable Ultralytics release. YOLO11 introduces architectural refinements over YOLOv8, including an updated C3k2 feature extraction module and an improved classification head with enhanced feature aggregation pathways. It is evaluated under conditions identical to YOLOv8n to allow direct architectural comparison.

E. Transfer Learning Strategy

All three models were initialized with weights pre-trained on ImageNet-1K, leveraging the rich hierarchical visual feature representations learned from large-scale natural image data. Full fine-tuning was applied: all backbone layers were updated during training on the orange disease dataset, rather than freezing any layer group. This decision reflects the relatively small size of the training set (approximately 841 images after the validation split). Prior empirical work on agricultural image classification with small domain-specific datasets has consistently shown that full fine-tuning produces better transfer outcomes than partial layer freezing, as the domain gap between natural ImageNet images and close-up agricultural fruit imagery is large enough that full feature adaptation is beneficial [1], [2]. Early stopping based on top-1 validation accuracy was applied during training to prevent overfitting.

F. Training Configuration

The training configuration below was applied consistently across all three models, with the explicitly noted exception of epoch count for the enhanced YOLOv8m. The SGD optimizer with momentum was preferred over adaptive optimizers (e.g., Adam) based on established YOLO training practice: SGD with cosine learning-rate decay has been shown to generalize better on fine-tuning tasks at moderate dataset sizes. Cosine annealing reduces the learning rate smoothly from its initial value toward zero over the training horizon, preventing overshooting during the final stages of fine-tuning. The enhanced YOLOv8m was trained for five additional epochs to accommodate its greater parameter count and slower initial convergence.

TABLE II. TRAINING HYPERPARAMETERS APPLIED ACROSS ALL MODELS

Hyperparameter	Value
Platform	Google Colab (Tesla T4 GPU, 14 GB VRAM)
Training Epochs	20 (YOLOv8n, YOLO11n); 25 (YOLOv8m Enhanced)
Batch Size	32
Input Resolution	224 × 224 pixels
Optimizer	SGD (momentum = 0.937)

G. Evaluation Protocol

Model performance was assessed exclusively on the held-out test set of 99 images, which was not used at any stage of training or model selection. The following metrics were calculated per class and aggregated: Top-1 Accuracy, the proportion of test images for which the highest-confidence predicted class matched the ground truth label; Precision (per class), the proportion of images predicted as a given class that genuinely belonged to that class; Recall (per class), the proportion of images actually belonging to a given class that were correctly identified as such; and F1-Score (per class), the harmonic mean of precision and recall. In addition to per-class metrics, macro-averaged and weighted-average aggregates are reported. Confusion matrices were computed for all three models to visualize inter-class error patterns, with particular attention to confusion between the visually similar Blackspot and Canker categories, which represent the most challenging classification boundary in this dataset.

V. RESULTS AND DISCUSSION

A. Overall Classification Performance

Table III summarizes the top-1 accuracy, macro precision, macro recall, and macro F1-Score for all three models on the held-out test set of 99 images. Precision, recall, and F1-Score are macro-averaged across the four defect classes, giving equal weight to each category regardless of its support in the test set.

TABLE III. OVERALL CLASSIFICATION PERFORMANCE ON THE TEST SET

Model	Top-1 Acc.	Macro Prec.	Macro Rec.	Macro F1
YOLOv8n (Baseline)	93.94%	0.93	0.94	0.93
YOLO11n	96.97%	0.97	0.97	0.97
YOLOv8m Enhanced	97.98%	0.98	0.98	0.98

The enhanced YOLOv8m achieved the highest test accuracy at 97.98%, followed by YOLO11n at 96.97% and YOLOv8n at 93.94%. The improvement from YOLOv8n to YOLOv8m Enhanced spans approximately four percentage points, which is meaningful given the small size of the test set and demonstrates the practical value of targeted augmentation for compact agricultural datasets. YOLO11n sits between the two YOLOv8 models, demonstrating that architectural improvements in the latest YOLO generation provide genuine performance gains even without changes to the training augmentation strategy.

B. Per-Class Performance

Table IV presents per-class F1-Scores for each model, disaggregated by defect category. The Support column reports the number of test images available for each class, allowing readers to judge the reliability of each per-class estimate.

TABLE IV. PER-CLASS F1-SCORE COMPARISON ACROSS ALL THREE MODELS

Class	YOLOv8n F1	YOLO11n F1	YOLOv8m Enh. F1	Support
Blackspot	0.88	0.93	0.95	22
Canker	0.87	0.93	0.95	22
Fresh	0.98	1.00	1.00	33
Greening	1.00	1.00	1.00	22

Across all three models, the Fresh and Greening classes were classified with high reliability, with all models achieving perfect or near-perfect F1-Scores on these categories. This is consistent with the expectation that Fresh oranges and Greening-affected fruit present visually distinctive characteristics that are straightforwardly separable from one another and from the two fungal/bacterial defect classes. The Blackspot and Canker classes presented greater classification challenges across all models. Both defects share overlapping visual characteristics—irregular surface textures and localized coloration changes—that make them inherently harder to distinguish. YOLOv8n recorded F1-Scores of 0.88 and 0.87 for Blackspot and Canker respectively. YOLO11n improved these to 0.93 for both classes through architectural refinement alone. The enhanced YOLOv8m reached 0.95 for both categories, confirming that the label smoothing and extended augmentation components provided measurable additional benefit on the most challenging classification boundary in this dataset. The confusion matrices in Fig. 3 illustrate the per-class prediction behavior of all three models across the four defect categories. Each row represents the actual class label, while each column represents the predicted class, allowing misclassification patterns to be identified at a glance. Across all three models, the Fresh and Greening classes are classified with near-perfect accuracy, reflecting the visual distinctiveness of these categories.

C. Comparative Analysis and Discussion

The results reveal a clear performance hierarchy consistent with the respective architectural capacities and training configurations of the three models. YOLOv8n, as the smallest model trained with default augmentation, establishes a solid baseline but struggles most with the Blackspot–Canker boundary. YOLO11n, despite sharing a similar nano-scale parameter budget, achieves meaningfully higher accuracy through refined feature aggregation architecture, narrowing the gap between baseline and enhanced performance without any additional preprocessing overhead. YOLOv8m Enhanced achieves the best overall results, though it is worth noting that this reflects the combined effect of larger model capacity and the extended augmentation and regularization pipeline. The independent contributions of model scale and augmentation strategy cannot be fully disentangled from this experimental design, a limitation that future ablation studies should address directly. Confusion matrix analysis confirms that the most frequent misclassification pattern across all three models involves confusion between the Blackspot and Canker categories. A controlled ablation, for example training a YOLOv8m variant with the default augmentation pipeline, or training YOLOv8n/YOLO11n with the extended pipeline, would be required to isolate the individual contributions of model capacity and augmentation strategy. Such an ablation was outside the scope of the present study and is identified as a priority for follow-up work; readers should therefore treat the reported YOLOv8m Enhanced results as reflecting the combined effect of both interventions rather than an architecture-only improvement.

The present evaluation was conducted on a curated dataset with relatively consistent imaging conditions, and the models' robustness to environmental variability characteristic of real packing-house settings, including inconsistent illumination, partial occlusion of the fruit surface by conveyor equipment or adjacent fruit, and sensor noise, was not directly assessed. Because the extended augmentation pipeline applied to the enhanced YOLOv8m already incorporates HSV colour jitter and random erasing, which partially emulate illumination variation and partial occlusion respectively, the model may exhibit some inherent robustness to these factors; however, this remains an assumption rather than an empirically validated claim. Future work should evaluate all three models on images captured under varying illumination, partial occlusion, and sensor-noise conditions to quantify robustness prior to field deployment.

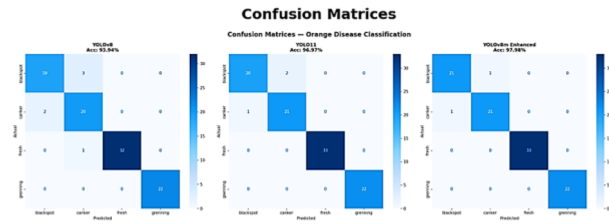

Fig. 3. Confusion matrices for YOLOv8n, YOLO11n, and YOLOv8m Enhanced. Rows indicate the true class label and columns indicate the predicted class; diagonal cells represent correct classifications, and off-diagonal cells highlight the Blackspot-Canker confusion discussed in Section V-B.

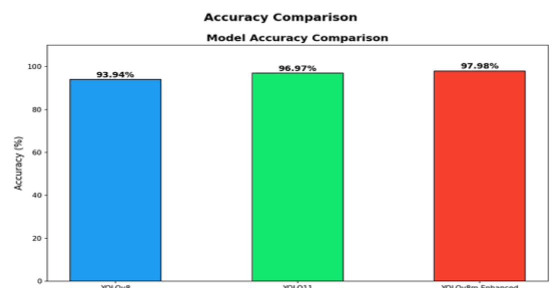

Fig. 4. Training and validation accuracy curves for YOLOv8n, YOLO11n, and YOLOv8m Enhanced. Solid and dashed lines denote training and validation accuracy, respectively; the x-axis represents training epoch and the y-axis represents top-1 accuracy.

Fig. 4 displays the top-1 accuracy curves recorded over the training epochs for all three models on both the training and validation sets. The curves illustrate how each model converged during fine-tuning and help assess whether overfitting occurred. YOLOv8n and YOLO11n, trained for 20 epochs, show relatively smooth convergence, with validation accuracy stabilizing in the later epochs. The enhanced YOLOv8m, trained for 25 epochs, benefits from the cosine annealing schedule and the additional regularization introduced by dropout and label smoothing, which keeps the gap between training and validation accuracy narrow throughout training, a sign that the model generalized well rather than memorizing training examples. Any divergence between the training and validation curves in the earlier epochs reflects the expected warm-up phase before the learning-rate schedule begins to take effect.

D. Qualitative Observations

Sample inference results from the YOLOv8m Enhanced model on test set images are consistent with the quantitative findings. The model assigns high-confidence predictions, typically above 0.90, to Fresh and Greening images, while Blackspot and Canker predictions range from 0.75 to 0.95 depending on defect severity and surface visibility. Misclassified images predominantly fall into two groups: early-stage defects where surface expression is minimal, and images where natural peel variation partially obscures the diagnostic features of the defect. Both failure modes point toward directions for future work, including dataset expansion with more examples of early-stage defects and imaging protocol standardization.

Fig. 5 presents a grouped bar chart comparing the per-class F1-Scores achieved by all three models across the four defect categories. The chart makes the performance hierarchy immediately visible: Fresh and Greening consistently reach F1-Scores at or near 1.00 across all models, confirming that these classes pose little classification difficulty regardless of model architecture. The Blackspot and Canker bars show a clear upward trend from YOLOv8n through YOLO11n to YOLOv8m Enhanced. Fig. 6 shows a selection of test images alongside the class label and confidence score predicted by the YOLOv8m Enhanced model. Correctly classified examples are shown for all four categories, illustrating the model's ability to distinguish between Fresh fruit, Greening discoloration, and the surface lesions characteristic of Blackspot and Canker. Confidence scores for Fresh and Greening predictions consistently exceed 0.90, reflecting the model's high certainty on visually unambiguous cases. Blackspot and Canker predictions generally fall in the 0.75–0.95 range, with lower confidence observed on early-stage defects where surface expression is minimal.

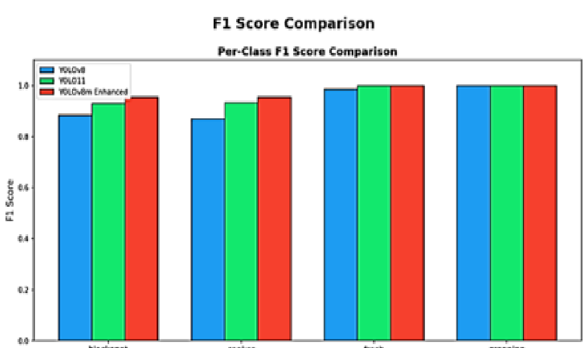

Fig. 5. Per-class F1-Score comparison across YOLOv8n, YOLO11n, and YOLOv8m Enhanced. Bars are grouped by defect class, with one bar per model per class, allowing direct visual comparison of the accuracy gains attributable to architecture (YOLOv8n vs. YOLO11n) and to the extended augmentation pipeline (YOLO11n vs. YOLOv8m Enhanced).

Fig. 7 plots the training loss recorded at each epoch for YOLO11n and YOLOv8m Enhanced, allowing a direct comparison of how the two models learned over their respective training runs. Both curves follow the expected downward trajectory, with steeper loss reduction in the early epochs as the pre-trained weights adapt to the orange disease dataset, followed by a more gradual decline as the models approach convergence.

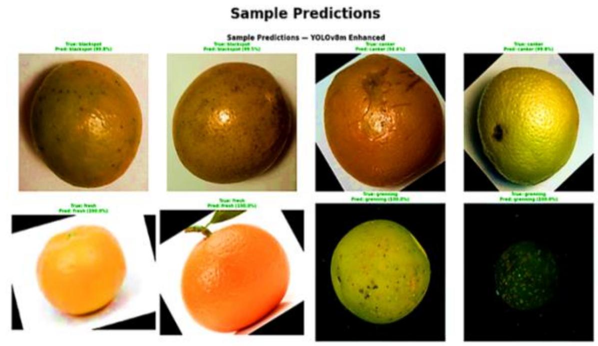

Fig. 6. Sample predictions from the YOLOv8m Enhanced model. Each panel shows the ground-truth class label alongside the YOLOv8m Enhanced model's predicted class and softmax confidence score.

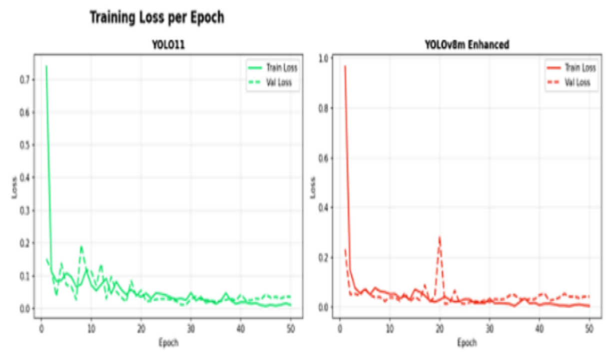

Fig. 7. Training loss curves for YOLO11n and YOLOv8m Enhanced. The x-axis represents training epoch and the y-axis represents training loss; both models use a cosine-annealing learning-rate schedule.

YOLO11n, trained for 20 epochs, shows a smoother and faster initial descent, reflecting the efficiency of its refined classification architecture. YOLOv8m Enhanced, trained for 25 epochs, exhibits a slightly slower early convergence due to the additional regularization introduced by dropout and label smoothing, but ultimately achieves a lower final training loss consistent with its superior test-set performance. The cosine annealing learning-rate schedule is visible in both curves as a smooth, continuous reduction in the rate of loss decrease toward the final epochs, preventing overshooting and supporting stable convergence.

VI. CONCLUSION

This paper presented a comparative evaluation of three YOLO-family classification models—YOLOv8n, YOLO11n, and an enhanced YOLOv8m—for automated orange fruit defect classification across four categories: Blackspot, Canker, Fresh, and Greening. All models were trained and evaluated under controlled and identical conditions on a publicly available orange disease dataset comprising 991 training images and 99 test images. All three models achieved strong classification performance. Top-1 accuracies ranged from 93.94% for YOLOv8n to 97.98% for the enhanced YOLOv8m. YOLO11n, at 96.97%, confirmed that architectural advances in the latest YOLO generation deliver meaningful improvements even in the absence of additional preprocessing. The enhanced YOLOv8m produced the strongest per-class F1-Scores, particularly on the visually challenging Blackspot and Canker categories, validating the effectiveness of the extended augmentation and regularization pipeline for small, class-imbalanced datasets. Taken together, these findings confirm the practical viability of YOLO-based classification models for agricultural defect grading, even in scenarios characterized by limited training data, class imbalance, and visual similarity between defect categories.

Several limitations should be considered when interpreting these results. First, the dataset is comparatively small (1,090 images in total, with only 99 test images), and although the validation split and fixed random seed support reproducibility, performance estimates on such a small test partition carry non-trivial sampling variance; confidence intervals or repeated-split evaluation would strengthen future reporting. Second, because the enhanced YOLOv8m combines a larger backbone with an extended augmentation and regularization pipeline, the reported performance gain cannot be attributed to architecture alone; a controlled ablation isolating these two factors is left for future work. Third, computational-complexity metrics, including training time, inference latency, and model size, together with formal statistical-significance testing across models, were outside the scope of the present study and should be reported in follow-up work to support practical deployment decisions. Finally, all models were trained and evaluated on a single curated dataset with relatively controlled imaging conditions; validation on larger, more diverse datasets collected under real-world packing-house conditions, including variable illumination, occlusion, and sensor noise, is recommended before these findings are generalized to operational deployment.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest to report regarding the present study.

AUTHOR CONTRIBUTIONS

Conceptualization, M.D.B., S.H., S.A. and H.B.; methodology, M.D.B., S.H., S.A., and S.A.; software, A.K., M.H.H. and I.K.; validation, M.D.B., S.H., S.A., H.B., M.H.H. and I.K.; writing—original draft preparation, M.D.B., S.H., S.A.; writing—review and editing, A.K., I.K., H.B. and M.H.H.

FUNDING STATEMENT

This research received no external funding.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

Data is available on reasonable request.

REFERENCES

[1] N. Butt, M. M. Iqbal, S. Ramzan, A. Raza, L. Abualigah, N. L. Fitriyani, Y. Gu, and M. Syafrudin, “Citrus diseases detection using innovative deep learning approach and Hybrid Meta-Heuristic,” PLoS ONE, vol. 20, no. 1, p. e0316081, 2025.

[2] F. Zhang, X. Jin, G. Lin, J. Jiang, M. Wang, S. An, J. Hu, and Q. Lyu, “Hybrid attention network for citrus disease identification,” Comput. Electron. Agric., vol. 220, p. 108907, 2024.

[3] A. Goyal and K. Lakhwani, “Integrating advanced deep learning techniques for enhanced detection and classification of citrus leaf and fruit diseases,” Sci. Rep., vol. 15, pp. 1–23, 2025.

[4] R. Mai, X. He, Y. Li, J. Ren, X. Gong, and G. Fan, “Advances in the machine learning construction process for citrus quality and disease detection,” Trends Food Sci. Technol., 2025.

[5] K. Peng, W. Ma, J. Lu, Z. Tian, and Z. Yang, “Application of machine vision technology in citrus production,” Appl. Sci., vol. 13, no. 16, p. 9334, 2023.

[6] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, “A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS,” Mach. Learn. Knowl. Extr., vol. 5, no. 4, pp. 1680–1716, 2023.

[7] R. Sapkota, M. F. Calero, R. Qureshi, C. Badgujar, U. Nepal, A. Poulose, P. Zeno, U. B. P. Vaddevolu, S. Khan, M. Shoman, H. Yan, and M. Karkee, “YOLO advances to its genesis: A decadal and comprehensive review of the You Only Look Once (YOLO) series,” Artif. Intell. Rev., 2025.

[8] Z. Zheng et al., “Improved YOLOv8-based algorithm for citrus leaf disease detection,” IEEE Access, 2025.

[9] H. Zhu, D. Wang, Y. Wei, P. Wang, and M. Su, “YOLOV8-CMS: A high-accuracy deep learning model for automated citrus leaf disease classification and grading,” Plant Methods, vol. 21, no. 1, p. 88, 2025.

[10] W. Xie, F. Feng, and H. Zhang, “A detection algorithm for citrus Huanglongbing disease based on an improved YOLOv8n (YOLO-EAF),” Sensors, vol. 24, no. 14, p. 4448, 2024.

[11] X. Guo, X. Wang, W. Zhu, S. X. Yang, L. Song, P. Li, and Q. Li, “Citrus disease detection based on dilated reparam feature enhancement and shared parameter head (YOLOv8n-DE),” Sensors, vol. 25, no. 7, p. 1971, 2025.

[12] H. Wang, Y. Wang, J. Liu, and J. Chang, “Research on citrus leaf disease detection based on improved SEDS-YOLOv8 model,” Int. Sci. Tech. Econ. Res., vol. 3, no. 1, 2025.

[13] T. Hong Son, D. Tran, T. Luong, D. Le, and H. Le, “A deep learning approach for real-time detection and evaluation of oranges on mobile devices,” 2025.

[14] H. Joshi, “Edge-AI for agriculture: Lightweight vision models for disease detection in resource-limited settings,” arXiv:2412.18635, 2024.

[15] R. Sapkota and M. Karkee, “Comparing YOLOv11 and YOLOv8 for instance segmentation of occluded and non-occluded immature green fruits in complex orchard environments,” arXiv:2410.19869, 2025.

[16] “Deep learning framework for fruit counting and yield mapping in tart cherry using YOLOv8 and YOLO11,” Comput. Electron. Agric., 2025.

[17] R. Sapkota, Z. Meng, M. Churuvija, X. Du, Z. Ma, and M. Karkee, “Comprehensive performance evaluation of YOLOv12, YOLO11, YOLOv10, YOLOv9 and YOLOv8 on detecting and counting fruitlet in complex orchard environments,” arXiv:2407.12040, 2024.

[18] S. Sifath, S. Saha, and S. Jahan, “Automatic classification of orange fruit diseases using deep neural network model,” J. Agric. Food Res., vol. 7, p. 102712, 2026.

[19] P. Kaushik and P. Sharma, “Deep learning-based detection of orange diseases using MobileNetV2,” in Proc. 13th Int. Conf. System Modeling & Advancement in Research Trends (SMART), 2024, pp. 23–27.

[20] K. N. Arifin, S. A. Rupa, M. M. Anwar, and I. Jahan, “Lemon and orange disease classification using CNN-extracted features and machine learning classifier,” arXiv:2408.14206, 2024.

[21] L. Kujur, V. Gupta, and A. Singhal, “A hybrid multi-optimizer approach using CNN and GB for accurate prediction of citrus fruit diseases,” Discover Appl. Sci., vol. 7, no. 3, p. 151, 2025.

[22] S. Srinivasan, L. Somasundharam, S. Rajendran, V. P. Singh, S. K. Mathivanan, and U. Moorthy, “DBA-ViNet: an effective deep learning framework for fruit disease detection and classification using explainable AI,” BMC Plant Biol., vol. 25, p. 7015, 2025.

[23] P. Dhiman, A. Kaur, V. R. Balasaraswathi, Y. Gulzar, A. A. Alwan, and Y. Hamid, “Image acquisition, preprocessing and classification of citrus fruit diseases: A systematic literature review,” Sustainability, vol. 15, no. 12, p. 9643, 2023.

[24] H. Zhu, L. Ren, M. Yang, Y. Zhang, C. Zhang, J. Cui, and J. Wu, "A systematic review of deep learning-based object detection methods for crop grading," CABI Agric. Biosci., vol. 6, no. 1, Art. 60, 2025.

[25] Y. Alhwaiti, M. Khan, M. Asim, M. Hameed, and M. Alruwaili, "Leveraging YOLO deep learning models to enhance plant disease identification," Sci. Rep., vol. 15, Art. 55410, 2025.

[26] Y. Miao, W. Meng, and X. Zhou, "SerpensGate-YOLOv8: An enhanced YOLOv8 model for accurate plant disease detection," Front. Plant Sci., vol. 15, p. 1514832, 2024.

[27] "YOLOv8-G: An improved YOLOv8 model for major disease detection in dragon fruit stems," Sensors, vol. 24, no. 15, p. 5034, 2024.

[28] W. Gómez-Flores, J. J. Garza-Saldána, and S. E. Varela-Fuentes, "CitrusUAT: A dataset of orange citrus sinensis leaves for abnormality detection using image analysis techniques," Data Brief, vol. 52, p. 109908, 2024.

[29] J. Silva, "Orange Diseases Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/jonathansilva2020/orange-diseases