

THREE ROADS TO DIAGNOSIS: BENCHMARKING CONVNEXT-TINY, DENSENET121, AND YOLOV8N-CLS FOR ROBUST TOMATO LEAF DISEASE RECOGNITION

Urwah Ambreen^{*1}, Asma Ahmed¹, Ali Husnain², Sahrish Khan³, Aqsa Ejaz⁴,
Anam Zahoor⁵, Muhammad Waqas Haider⁶, Muhammad Usman⁷, Faizan Saleem⁸,
Muhammad Saqib⁹

^{*1,3,4,5,6,7,8,9} Faculty of Information Technology, The University of Lahore, Sargodha Campus, Sargodha 40100, Pakistan

¹urwahambreen234@gmail.com, ¹asmaahmad13822@gmail.com, ²alihanain90@yahoo.com,
³sahrishkhan1215@gmail.com, ⁴zulqarnainaqs3@gmail.com, ⁵malikanam06@gmail.com, ⁶whaider491@gmail.com,
⁷usmanituos1@gmail.com, ⁸faizan.saleem@cs.uol.edu.pk, ⁹saqibaliriaz888@gmail.com,

DOI: <https://doi.org/10.5281/zenodo.23159677>

Keywords

Tomato leaf disease; convnext;
densenet; yolov8; transfer
learning; precision agriculture;
domain generalisation

Article History

Received: 25 June 2026

Accepted: 16 September 2026

Published: 05 October 2026

Copyright @Author

Corresponding Author: *

Urwah Ambreen

Abstract

Tomato leaf disease is still diagnosed manually across most producing regions, a process too slow and subjective to scale to modern field sizes. This paper narrows the question asked by a companion seven-architecture benchmark to a sharper, three-way comparison: how do a lightweight single-stage detector repurposed for classification (YOLOv8n-cls), a densely-connected convolutional network (DenseNet121), and a modernised pure-convolutional backbone (ConvNeXt-Tiny) compare, not only on curated laboratory imagery but on their ability to survive the transition to field-captured photography? Using a class-balanced, 23,727-image, ten-class tomato leaf dataset, all three architectures were fine-tuned under an identical protocol and evaluated on a held-out 1,500-image in-distribution test split. YOLOv8n-cls and DenseNet121 tied for the best controlled-condition result at 96.20% accuracy and 96.19% weighted F1-score, with ConvNeXt-Tiny close behind at 96.13% accuracy. The same three architectures were then independently retrained and evaluated on a five-class, field-captured subset of the CCMT dataset to probe generalisation. Here the ranking inverted: ConvNeXt-Tiny, the weakest of the three under laboratory conditions, generalised best by a wide margin (85.32% field accuracy, 85.01% F1-score), followed by DenseNet121 (81.88%, 81.62%), while YOLOv8n-cls – despite tying for first place on curated imagery – generalised worst of the three (77.64%, 76.60%). The gap between laboratory and field accuracy therefore ranged from 10.81 points (ConvNeXt-Tiny) to 18.56 points (YOLOv8n-cls), a reversal that a laboratory-only comparison of these three models would never have revealed. We argue that architecture family, not parameter count or laboratory accuracy, is the more reliable signal of field robustness among the three backbones studied, and we discuss what this means for choosing a deployment architecture in practice.

INTRODUCTION

Tomato (*Solanum lycopersicum*) is grown across nearly every agro-climatic zone in the world and remains one of the highest-value vegetable crops for smallholder and commercial farmers alike. Its yield, however, is chronically eroded by leaf-borne diseases bacterial spot, early and late blight, leaf mould, septoria leaf spot, spider-mite infestation, target spot, mosaic virus and the yellow-leaf-curl virus among the most common – several of which can reduce marketable yield by a third or more if left undiagnosed for even a short window [10], [11], [12]. In most producing regions, diagnosis is still performed visually by an agronomist or the grower, a process that is slow, subjective and difficult to scale to the size of a modern field.

Deep-learning classifiers routinely exceed 95% accuracy on curated tomato-leaf benchmarks such as PlantVillage [3], but that figure is measured against a plain, uniformly-lit background that has little in common with a phone photograph taken in an actual field. A recent benchmarking study, TLD46k [1], made this gap explicit, showing that YOLO11 classifiers trained on one source domain lost 30–75 percentage points of accuracy when evaluated on another. The present paper follows the same line of inquiry but narrows the architectural scope to three representative, widely-used backbones – one lightweight detector-derived classifier (YOLOv8n-cls) and two ImageNet-pretrained CNN families (DenseNet121 and ConvNeXt-Tiny) – so that the trade-off between laboratory accuracy and field robustness can be examined closely for exactly the three architectures a practitioner is most likely to be choosing between when parameter budget, not benchmark novelty, is the constraint.

This paper contributes: (i) a head-to-head, matched-protocol comparison of ConvNeXt-Tiny, DenseNet121 and YOLOv8n-cls on a curated, class-balanced, 23,727-image tomato leaf dataset; (ii) an independent field-condition generalisation test of the same three architectures on a held-out subset of the field-captured CCMT dataset; and (iii) evidence that, among these three specific architectures, the model that wins under laboratory conditions is not the model best suited to field deployment.

ii. Related Work

Early large-scale tomato-disease datasets were dominated by PlantVillage [3], which supplies over 50,000 leaf images across 14 crop species photographed against a plain background under controlled lighting; several audits have since shown that this uniformity lets even shallow classifiers reach near-perfect accuracy while learning little about real foliage variation [1], [4]. CCMT [5] sits at the other end of the spectrum: collected under semi-controlled field photography for cashew, cassava, maize and tomato, it offers substantially more visual diversity while remaining large enough to train a classifier from scratch, and is used here as the field-condition test bed.

DenseNet121 [6] remains a strong, widely-used transfer-learning baseline for plant-disease classification because its dense inter-layer connectivity encourages feature reuse and its ImageNet pretraining transfers well to small, domain-specific datasets; unlike the additive skip connections popularised by ResNet [16], DenseNet concatenates feature maps across layers, a design choice reported to improve gradient flow and parameter efficiency. ConvNeXt [7] modernises the same convolutional recipe with design choices borrowed from the transformer literature [17] – larger kernel sizes, an inverted bottleneck and layer normalisation – while keeping a pure-convolutional backbone; it has not, to our knowledge, previously been benchmarked against DenseNet and a YOLO-derived classifier specifically on tomato disease imagery under a matched field-generalisation protocol. YOLOv8 [8] belongs to a separate lineage: originally a single-stage object detector [19] descended from the anchor-based ideas of Faster R-CNN [13] and SSD [14], its classification head repurposes the same efficient backbone for whole-image labelling, trading some of the representational capacity of a dedicated CNN for a substantially smaller parameter count.

The work closest to ours is TLD46k [1], which cross-validates YOLO11 classifiers across merged PlantVillage, CCMT and PlantDoc [9] subsets and finds that no single source generalises well to the others. That study, by design, restricts itself to one architecture family. The present paper asks a complementary question on a fixed, smaller set of three architectures: given equal training data and

protocol, which of a lightweight detector-derived classifier, a dense CNN, and a modernised convolutional network best survives the move from lab to field – and does the laboratory leaderboard predict the field leaderboard at all?

Iii. Materials And Methods

A. Overall Pipeline

Fig. 1 summarises the end-to-end workflow used in this study: the curated 23,727-image tomato collection is cleaned and resized, split into class-balanced train, validation and test partitions, used to independently fine-tune the three candidate architectures, and finally evaluated both in-distribution and on the held-out, field-captured CCMT subset.

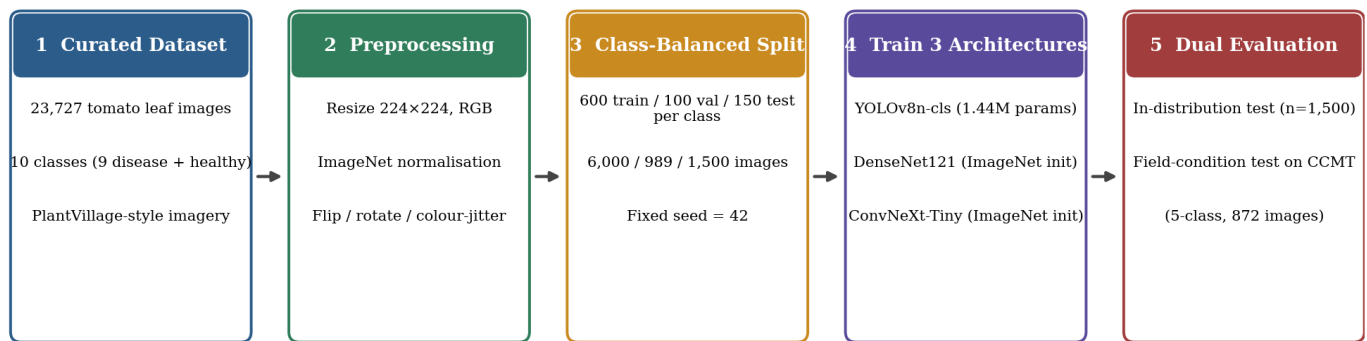


Fig. 1. End-to-end methodology pipeline for the three-architecture benchmark, from curated imagery to dual-condition evaluation.

B. Dataset

The primary experiment uses a curated, tomato-only collection of 23,727 leaf images spanning ten categories – nine disease classes and a healthy control – assembled from PlantVillage-style photography (Table I, Fig. 2). The raw class distribution is imbalanced by close to a factor of five (Yellow Leaf Curl Virus, 5,944 images, versus Mosaic Virus, 1,163 images), a ratio typical of naturally-collected disease imagery and retained here rather than artificially flattened. For computational tractability on a single T4 GPU, training itself was carried out on a class-balanced sub-sample – 600 training, ~100 validation

and 150 test images per class (6,000 / 989 / 1,500 images respectively).

A second, independent dataset was used exclusively to probe field-condition generalisation: the tomato subset of CCMT [5], restricted to the five classes common across its raw collection (healthy, leaf blight, leaf curl, septoria leaf spot and verticillium wilt) and split into 4,042 training, 867 validation and 872 test images. Because CCMT images are captured directly in cultivation settings rather than against a plain backdrop, this subset stands in for the field condition discussed in Section IV-B.

Table I. Class-Wise Split Of The Curated 10-Class Tomato Dataset

Class	Train	Val.	Test	Total
Bacterial Spot	1,602	178	445	2,225
Early Blight	882	98	244	1,224
Late Blight	1,451	161	402	2,014

Class	Train	Val.	Test	Total
Leaf Mold	889	98	246	1,233
Mosaic Virus	838	93	232	1,163
Septoria Leaf Spot	3,358	373	932	4,663
Spider Mites	1,220	135	338	1,693
Target Spot	1,033	114	286	1,433
Yellow Leaf Curl Virus	4,281	475	1,188	5,944
Healthy	1,538	170	427	2,135
Total (72.0/8.0/20.0%)	17,092	1,895	4,740	23,727

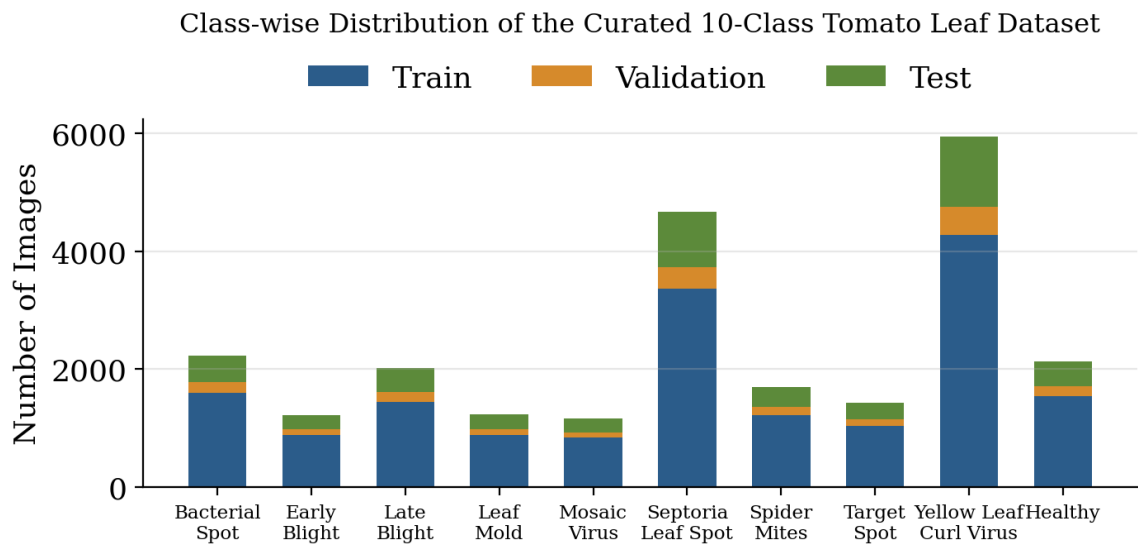


Fig. 2. Class-wise distribution of the curated 10-class tomato leaf dataset (train / validation / test).

C. Preprocessing

All images were converted to RGB, resized to 224×224 pixels and normalised with ImageNet channel statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). During training, the CNN pipeline (DenseNet121, ConvNeXt-Tiny) additionally applied random horizontal flipping, random rotation ($\pm 15^\circ$) and colour-jitter (brightness and contrast, $\pm 20\%$) to reduce over-fitting; YOLOv8n-cls used Ultralytics' built-in classification augmentation pipeline (mosaic, flips and HSV jitter) at default settings.

D. Model Architectures

Three classifiers were trained and compared. YOLOv8n-cls was taken from Ultralytics' pretrained classification checkpoints; at 1.44 million parameters it is roughly an order of magnitude smaller than a

fine-tuned DenseNet121, making it a natural candidate for on-device or edge deployment. DenseNet121 [6] and ConvNeXt-Tiny [7] were initialised from ImageNet-1k weights and fine-tuned end-to-end with a task-specific classification head (a dropout-linear-ReLU-dropout-linear stack for DenseNet121).

E. Training Configuration

All three models were trained under a matched protocol so that reported differences reflect architecture rather than schedule: 10 epochs, batch size 32, AdamW optimisation [15], early-stopping patience of 4 epochs and a fixed random seed (42) for reproducibility (Table II).

Table Ii. Training Configuration

Hyper-parameter	Value
Epochs	10
Early-stopping patience	4
Batch size	32
Learning rate	1e-3 (CNN); auto (YOLO)
Optimizer	AdamW
Input resolution	224 × 224 px
Random seed	42
GPU	NVIDIA Tesla T4 (15.6 GB)

F. Evaluation Protocol

Performance is reported with accuracy, precision, recall and weighted F1-score. Two evaluation regimes are used: an in-distribution test on the balanced 1,500-image tomato test split (Section IV-A), and an independent field-condition test on the held-out CCMT split (Section IV-B), which approximates the lab-to-field domain shift highlighted by TLD46k [1].

Iv. Results And Discussion

A. Controlled-Condition Performance

Table III and Fig. 3 summarise the weighted-average metrics of the three classifiers on the balanced, 1,500-image, in-distribution test split. YOLOv8n-cls and DenseNet121 tie for first place at 96.20% accuracy and 96.19% weighted F1-score; ConvNeXt-Tiny is a close third at 96.13% accuracy and 96.11% F1-score. The spread across all three is only 0.07 percentage points, reinforcing that on curated imagery of this kind, architecture choice among reasonably deep, ImageNet-pretrained networks is a second-order effect.

Table Iii. Controlled-Condition Test Results (N = 1,500)

Rank	Model	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
1	YOLOv8n-cls	96.20	96.26	96.20	96.19
1	DenseNet121	96.20	96.23	96.20	96.19
3	ConvNeXt-Tiny	96.13	96.21	96.13	96.11

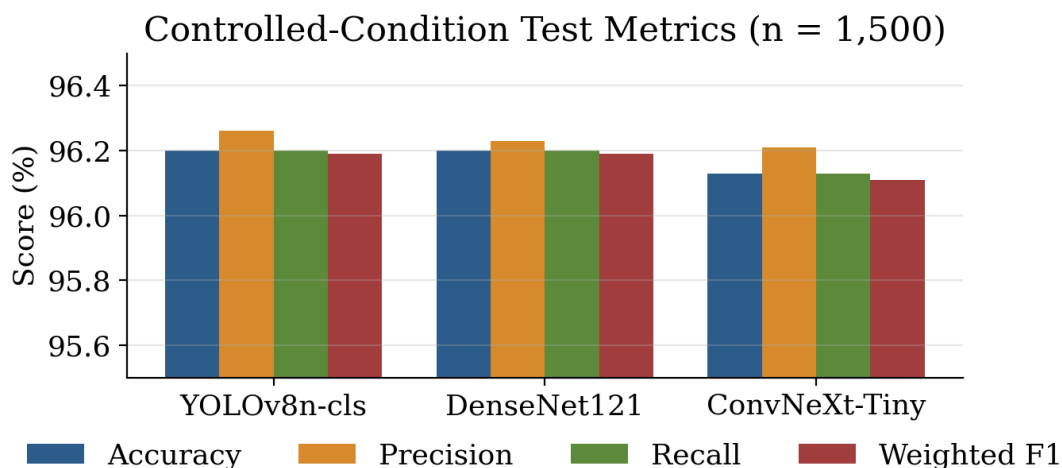


Fig. 3. Accuracy, precision, recall and weighted F1-score of the three classifiers on the controlled-condition test split.

Per-class inspection of the two best-tied models (reported in the companion seven-architecture study on the same dataset) shows both YOLOv8n-cls and DenseNet121 dip below 93% F1-score on two classes in particular: Septoria Leaf Spot (91.15% and 91.22% F1 respectively) and Tomato Yellow Leaf Curl Virus (92.73% and 92.57% respectively), with the Yellow Leaf Curl Virus error being recall-dominated rather than precision-dominated – i.e., both models under-detect this class more often than they over-predict it, despite it being the single largest class in the training set.

B. Field-Condition Generalisation Test

The generalisation test retrains and evaluates all three architectures on the five-class CCMT split described in Section III-B. Table IV and Fig. 4 report the resulting accuracy, F1-score and the generalisation gap relative to each model's own laboratory test accuracy. ConvNeXt-Tiny generalises best by a clear margin (85.32% accuracy, 85.01% F1-score, a gap of only 10.81 points), followed by DenseNet121 (81.88%, 81.62%, gap 14.32 points); YOLOv8n-cls, despite tying for the best laboratory result, generalises worst of the three (77.64%, 76.60%, gap 18.56 points).

Table IV. Field-Condition (Ccmt) Generalisation Results

Model	Lab Acc. (%)	Field Acc. (%)	Field F1 (%)	Gap (pts)
ConvNeXt-Tiny	96.13	85.32	85.01	10.81
DenseNet121	96.20	81.88	81.62	14.32
YOLOv8n-cls	96.20	77.64	76.60	18.56

Lab vs. Field-Condition Accuracy: ConvNeXt-Tiny, DenseNet121, YOLOv8n-cls

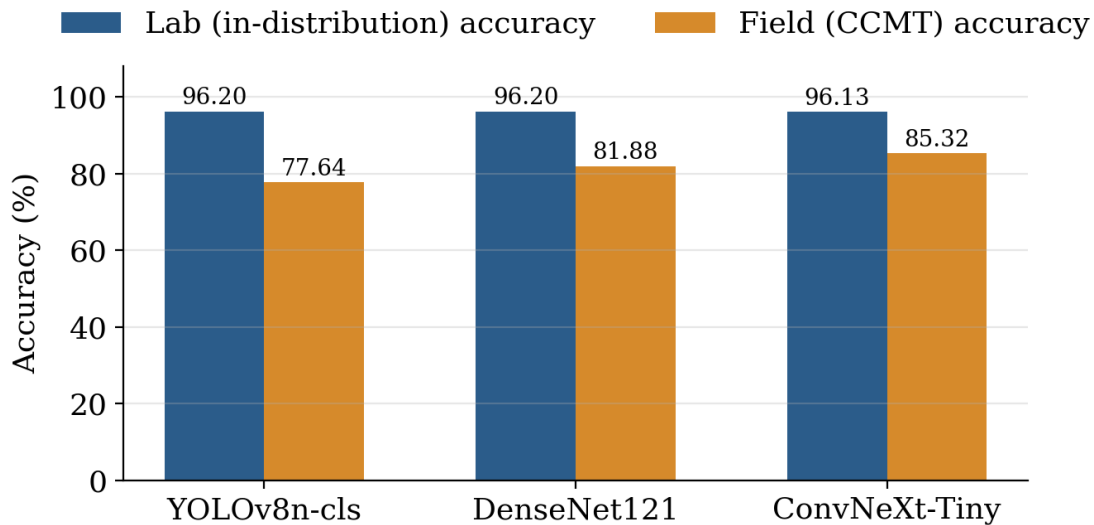


Fig. 4. Lab (in-distribution) versus field-condition (CCMT) accuracy for ConvNeXt-Tiny, DenseNet121 and YOLOv8n-cls.

This inversion – the two models that tied for first place under laboratory conditions swap to being the two worst generalisers, while the laboratory third-place model becomes the clear field-condition winner – is the central finding of this narrower, three-architecture comparison (Fig. 5). It mirrors, on an

independent and smaller set of models, the domain-shift problem documented for TLD46k [1], and it suggests that among ConvNeXt-Tiny, DenseNet121 and YOLOv8n-cls specifically, a practitioner choosing on the basis of laboratory accuracy alone would select the architecture least suited to field deployment.

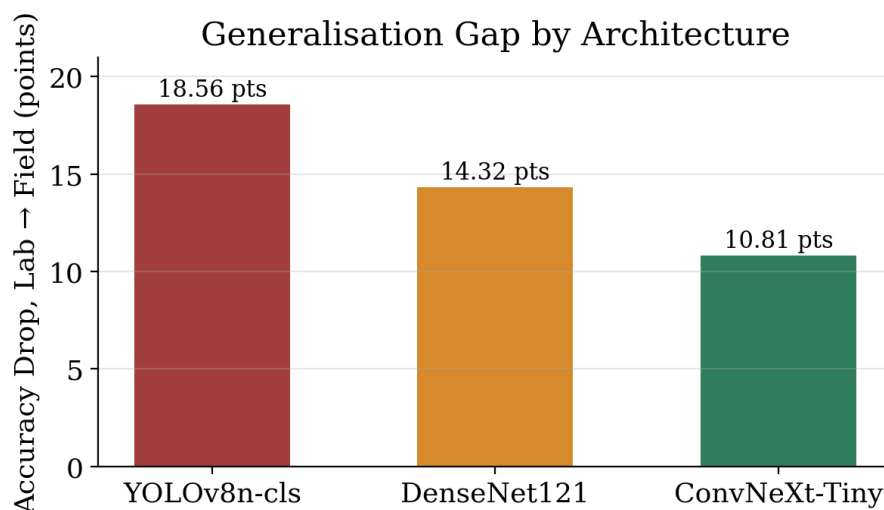


Fig. 5. Generalisation gap (laboratory test accuracy minus field accuracy) by architecture.

C. Why the Ranking Inverts

A plausible explanation lies in inductive bias rather than parameter count. ConvNeXt-Tiny's larger effective receptive field and transformer-inspired design choices (large-kernel depthwise convolutions, layer normalisation) may encourage it to rely on more holistic, less background-dependent cues than the smaller, detector-derived YOLOv8n-cls backbone, which was originally optimised for localisation on comparatively clean imagery. DenseNet121's dense connectivity sits between the two, both in parameter count and in generalisation behaviour. Confirming this explanation would require gradient-based attribution (e.g., Grad-CAM [18]) comparing what each network attends to on CCMT imagery, which we leave for future work.

V. Conclusion And Future Work

This paper compared three widely-used architectures – ConvNeXt-Tiny, DenseNet121 and YOLOv8n-cls – on a curated tomato leaf disease dataset and on an independent, field-captured generalisation test. Under laboratory conditions the three models were effectively tied, clustering within 0.07 percentage points of one another at approximately 96% accuracy. Under field conditions the ranking inverted entirely: ConvNeXt-Tiny, the weakest of the three in the laboratory, became the strongest generaliser, while YOLOv8n-cls, tied for the best laboratory result, generalised worst. For practitioners choosing among exactly these three architectures, this result argues for evaluating on field-representative imagery before deployment, rather than relying on a laboratory benchmark leaderboard.

Future work should extend this three-way comparison with gradient-based visual explanations (Grad-CAM [18]) to confirm the inductive-bias hypothesis in Section IV-C, and should evaluate whether merging PlantVillage-style and CCMT-style imagery during training narrows the generalisation gap for all three architectures, as suggested by the TLD46k benchmark [1].

Acknowledgment

The authors thank Muhammad Saqib for his supervision, technical guidance and continued feedback throughout the design, implementation and evaluation of this study.

REFERENCES

- [1] T. K. Pham, K. L. Nguyen, and T. H. Dinh, "TLD46k: A Benchmark for Tomato Leaf Disease Detection," in Proc. ICTA 2025, Lecture Notes in Networks and Systems, vol. 1831, Springer, 2026, pp. 482–490.
- [2] Y. Alhwaiti, M. Khan, M. Asim, M. H. Siddiqi, M. Ishaq, and M. Alruwaili, "Leveraging YOLO deep learning models to enhance plant disease identification," *Sci. Rep.*, vol. 15, no. 1, art. 7969, 2025.
- [3] D. P. Hughes and M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," *arXiv:1511.08060*, 2015.
- [4] M. A. Noyan, "Uncovering bias in the PlantVillage dataset," *arXiv:2206.04374*, 2022.
- [5] P. K. Mensah et al., "CCMT: Dataset for crop pest and disease detection," *Data in Brief*, vol. 49, art. 109306, 2023.
- [6] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE CVPR, 2017, pp. 4700–4708.
- [7] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in Proc. IEEE/CVF CVPR, 2022, pp. 11976–11986.
- [8] G. Jocher, A. Chaurasia, and J. Qiu, "YOLOv8," Ultralytics, 2023. [Online]. Available: <https://docs.ultralytics.com/models/yolov8/>
- [9] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, and N. Batra, "PlantDoc: a dataset for visual plant disease detection," in Proc. 7th ACM IKDD CoDS and 25th COMAD, 2020, pp. 249–253.
- [10] M. Ahmad, M. Abdullah, H. Moon, and D. Han, "Plant disease detection in imbalanced datasets using efficient convolutional neural networks with stepwise transfer learning," *IEEE Access*, vol. 9, pp. 140565–140580, 2021.
- [11] R. N. Strange and P. R. Scott, "Plant disease: a threat to global food security," *Annu. Rev. Phytopathol.*, vol. 43, no. 1, pp. 83–116, 2005.
- [12] E.-C. Oerke and H.-W. Dehne, "Safeguarding production—losses in major crops and the role of crop protection," *Crop Prot.*, vol. 23, no. 4, pp. 275–285, 2004.

- [13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NeurIPS*, 2015.
- [14] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. ECCV*, 2016, pp. 21–37.
- [15] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.
- [17] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
- [18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.
- [19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE CVPR*, 2016, pp. 779–788.
- [20] L. Goyal, C. M. Sharma, A. Singh, and P. K. Singh, "Leaf and spike wheat disease detection & classification using an improved deep convolutional architecture," *Informat. Med. Unlocked*, vol. 25, Jan. 2021, Art. no. 100642.
- [21] M. Turkoglu, B. Yanikoğlu, and D. Hanbay, "PlantDiseaseNet: Convolutional neural network ensemble for plant disease and pest detection," *Signal, Image Video Process.*, vol. 16, no. 2, pp. 301–309, Mar. 2022.
- [22] S. Nalini, N. Krishnaraj, T. Jayasankar, K. Vinothkumar, A. S. F. Britto, K. Subramaniam, and C. Bharatiraja, "Paddy leaf disease detection using an optimized deep neural network," *Comput., Mater. Continua*, vol. 68, no. 1, pp. 1117–1128, 2021.
- [23] S. F. Syed-Ab-Rahman, M. H. Hesamian, and M. Prasad, "Citrus disease detection and classification using end-to-end anchor-based deep learning model," *Int. J. Speech Technol.*, vol. 52, no. 1, pp. 927–938, Jan. 2022.
- [24] R. Mahum, H. Munir, Z.-U.-N. Mughal, M. Awais, F. S. Khan, M. Saqlain, S. Mahamad, and I. Tlili, "A novel framework for potato leaf disease detection using an efficient deep learning model," *Human Ecol. Risk Assessment*, Int. J., vol. 29, no. 2, pp. 303–326, Feb. 2023.
- [25] C. Madhurya and E. A. Jubilson, "YR2S: Efficient deep learning technique for detecting and classifying plant leaf diseases," *IEEE Access*, vol. 12, pp. 3790–3804, Jan. 2024, doi: 10.1109/ACCESS.2023.3343450.
- [26] M. P. Mathew and T. Y. Mahesh, "Leaf-based disease detection in bell pepper plant using YOLO v5," *Signal, Image Video Process.*, vol. 16, no. 3, pp. 841–847, Apr. 2022.
- [27] M. Mishra, P. Choudhury, and B. Pati, "Modified ride-NN optimizer for the IoT based plant disease detection," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 1, pp. 691–703, Jan. 2021.
- [28] S. Vallabhajosyula, V. Sistla, and V. K. K. Kolli, "Transfer learning-based deep ensemble neural network for plant leaf disease detection," *J. Plant Diseases Protection*, vol. 129, no. 3, pp. 545–558, Jun. 2022.
- [29] Y. Wang, H. Wang, and Z. Peng, "Rice diseases detection and classification using attention based neural network and Bayesian optimization," *Expert Syst. Appl.*, vol. 178, Sep. 2021, Art. no. 114770.
- [30] A. Oad et al., "Plant leaf disease detection using ensemble learning and explainable AI," *IEEE Access*, vol. 12, Oct. 2024, doi: 10.1109/ACCESS.2024.3484574.
- [31] E. A. Aldakheel, M. Zakariah, and A. H. Alabdall, "Detection and identification of plant leaf diseases using YOLOv4," *Front. Plant Sci.*, vol. 15, p. 1355941, Apr. 2024, doi: 10.3389/fpls.2024.1355941.
- [32] Y. M. A. Algani et al., "Leaf disease identification and classification using optimized deep learning," *Meas. Sens.*, vol. 25, p. 100643, 2023.