

Explainable transfer learning-based approach for Mango fruit damage detection and cause identification

1st Yonasi Safari*

Department of Computer Science
Mbarara University of Science and Technology
Mbarara, Uganda
syonasi@must.ac.ug

2nd Rose Nakibuule

Department of Computer Science
Makerere University
Kampala, Uganda
rose.nakibuule@mak.ac.ug

3rd Joyce Nabende, Rose Nakasi

Department of Computer Science
Makerere University
Kampala, Uganda
joyce.nabende@mak.ac.ug

Abstract—This study presents a deep learning approach to detect mango fruit damage and identify its underlying causes using one-stage and two-stage detectors. Unlike existing methods that focus mainly on detection, this approach integrates XAI techniques to provide interpretable, cause-level insights, and is evaluated on a dataset annotated with anthracnose, bacterial black spot, fruit fly, and other damage classes. RetinaNet achieved the best overall performance, attaining a precision of 0.77, recall of 0.75, and F1-score of 0.76. Per-class accuracy exceeded 0.80 in all categories, except for fruitfly, which remained more challenging to detect. XAI techniques that include Grad-CAM, Grad-CAM++, LIME, SHAP, and LRP were further used to highlight vital regions influencing model decisions. A quantitative evaluation of XAI performance was further conducted using faithfulness, robustness, complexity, localization, and randomization. LRP and SHAP results attained are indicative of more faithful and well-localized explanations, while Grad-CAM-based methods exhibited limitations in localization, with LIME producing highly complex explanations that result in reduced interpretability.

Index Terms—Mango orchard, Fruit damage detection, Fruit damage cause, Object detection, One-stage detectors, Two-stage detectors, Explainable AI (XAI)

I. INTRODUCTION

Mango is a lucrative crop globally and in Uganda, where horticulture plays a key role in economic growth. However, pests, diseases, and climate-related challenges which lead to many infections, causing significant fruit losses, thus reducing production, endangering orchards, and affecting farmers' income from growing, processing, and exporting mangoes [1]. The mango fruit industry accordingly depends on early detection of fruit damage and the underlying causes to minimise such losses [2]. Precise detection and assessment of mango damage severity are crucial for effective damage management, helping farmers apply timely interventions to reduce loss. To support this, numerous automated non-destructive methods using traditional machine learning (ML) techniques have been developed for tasks such as detection and classification of fruit defects. However, ML techniques rely on manual feature

extraction, which is challenging. Deep learning (DL) techniques are exceptionally good at learning complex patterns and representations from images, eliminating the need for manually engineered features. Both ML and DL based models are often seen as black boxes because their decision-making process is not transparent. This lack of interpretability reduces user trust, thus limiting their adoption in precision agriculture. This has called for the use of Explainable AI (XAI) techniques to provide clear, visual explanations of predictions, improving transparency and trust. Due to the wide range of infections affecting mango fruits [3], there is a growing need for reliable and automated systems for their detection and classification. However, most existing studies have primarily focused on post-harvest analysis while a few have been at the orchard level [4] using deep learning-based techniques. To enhance existing approaches, the development of hybrid models has been recommended to combine detection and classification networks into a unified framework to simplify training procedures [4].

A. Research Contribution

- 1) An orchard-level mango dataset annotated for damage regions and underlying causes.
- 2) A comparative analysis of one-stage and two-stage models for damage cause detection.
- 3) Integration of XAI techniques to interpret and validate deep learning predictions for multiple mango damage causes.

II. RELATED WORK ON DAMAGE DETECTION IN FRUITS

Aldiki et al. [5] used EfficientNet models (B0–B4) to detect apple diseases and pests, incorporating Grad-CAM for interpretability. Performance increased with model complexity, EfficientNet B4 attaining the best results with precision, recall, F1 score of 0.91, and accuracy of 92%. Similarly, Sireesha et al. [6] proposed an interpretable CNN+ViT+SVM ensemble model for detecting citrus disease, using Grad-CAM, Grad-CAM++, and LIME, attaining up to 99% accuracy, IoU and Dice of 1.00 respectively. Zhang et al. [7] also used a ResNet50 model to classify healthy and damaged fruits,

This research study was done under AdEMNEA, a NORAD-funded project through the NORHED2 program under grant agreement QZA-21/0159

attaining 92% accuracy, with Grad-CAM applied to enhance interpretability and transparency. Priya et al. [8] also proposed a feature fusion approach combining ResNet50 and MobileNetV2 with PCA and Gradient Boosting, achieving 95.7% accuracy on mango defect classification, and further employed Grad-CAM for visualization. Yunong et al. [9] proposed a CycleGAN and YOLOv3-Dense approach for detecting apple anthracnose in orchards, attaining 95.57% accuracy. Zhang et al. [4] then proposed a two-stage deep learning-based framework for canker, anthracnose, sunscald, greening, and melanose citrus diseases detection and classification, achieving 89.0% accuracy and 87.2% F1-score, with optimized YOLOv4 for detection and EfficientNet for classification outperforming other models. Ramazan et al. [10] employed ResNet-50, GoogleNet, VGG-16, and AlexNet, in conjunction with SGDM, RMSProp and Adam optimisers to classify healthy and Mediterranean fruitfly-infected citrus fruits, giving accuracies of up to 99.33%. Haque et al. [11] also developed an automated system using CNN to detect and classify anthracnose, fruit rot, and canker guava diseases in orchards and also used online images, attaining a classification accuracy of 95.61%. Generally, most existing studies focus on fruit damage segmentation, detection, and classification for specific fruit types [12], with limited work at the orchard level [13] and little emphasis on identifying the underlying causes of damage.

III. METHODOLOGY

1) *Dataset Collection and Annotation:* A dataset of 6,347 mature mango fruit images collected under natural orchard conditions at Delight Farm in northern Uganda was manually annotated into four damage categories: anthracnose, bacterial black spot, fruit fly damage, and others using entomologists. Polygon-based labeling using Roboflow was used, and the dataset was split into training, validation, and test sets in COCO format. Image preprocessing and augmentation techniques including flipping, rotation, scaling, cropping, brightness/contrast adjustment, and noise addition were used to improve model robustness.

A. Model selection and training

This study employed transfer learning-based one-stage and two-stage object detection models leveraging COCO-pretrained weights, with implementation training done in PyTorch and torchvision.

1) *Single-Stage Object Detection Models used:* Models employed here include YOLOv11 that improves detection performance through architectural enhancements such as the C2PSA module and the efficient C3k2 block [14], SSD (Single Shot MultiBox Detector) which predicts object classes and bounding boxes directly from multiple feature maps with different resolutions to naturally handle objects of various sizes, and RetinaNet that utilizes a focal loss function to address class imbalance during training [15].

2) *Two-Stage Object Detection Models:* Faster R-CNN that enhances object detection by replacing selective search algorithm with a Region Proposal Network that efficiently generates candidate object regions and bounding boxes for classification. Mask R-CNN extends Faster R-CNN by incorporating a mask prediction phase for pixel-level segmentation [15].

B. Explainable Artificial Intelligence (XAI)

XAI models enhance trust, transparency and interpretability of deep learning models, which are often seen as “black boxes” due to their opaque decision-making processes. Models employed include; Gradient-weighted Class Activation Mapping (Grad-CAM), Grad-CAM++, Local Interpretable Model-agnostic Explanations (LIME), SHAP and Layer-wise Relevance Propagation (LRP) [16] [17].

C. Implementation of the models

Experiments were conducted on a RunPod cloud environment using NVIDIA A40 GPU, with models trained using the AdamW optimizer, lr 0.001, image size 640 px and batch sizes adjusted based on model complexity. The DL models were evaluated using standard metrics such as precision, recall, and F1-score [18], while XAI methods used explanation quality metrics including:- Faithfulness that measures how accurately explanations reflect the model’s decision process, then complexity that assesses their interpretability. Localization evaluates how well explanations align with ground-truth regions or bounding boxes. Robustness measures the stability of explanations under small input changes, and randomization assesses how explanations change when model parameters or inputs are randomly perturbed [19], [20].

IV. EXPERIMENTAL RESULTS

A. Localisation, Precision, Recall, and F1 score

All models attained strong performance with mean IoU values above 0.84, indicating effective localization of mango damage regions. YOLOv11 performed best with a mean IoU of 0.903, followed by Mask R-CNN (0.865), RetinaNet (0.863), Faster R-CNN (0.846), and SSD (0.845). Further more all models demonstrated effective damage detection capability, with RetinaNet attaining the highest performance (precision: 0.77, recall: 0.75, F1-score: 0.76), likely due to its focal loss mechanism for handling class imbalance. Mask R-CNN also showed strong results (0.69, 0.69, 0.69), followed by SSD (0.67, 0.68, 0.67), while YOLOv11 (0.59, 0.59, 0.58) and Faster R-CNN (0.60, 0.59, 0.58) recorded comparatively lower performance. Overall, the results indicate that model architecture and feature learning strategies significantly influence detection performance.

B. The per-class accuracy

Per-class accuracy varied across damage categories: for anthracnose, YOLOv11 (0.54), SSD (0.70), RetinaNet (0.83), Faster R-CNN (0.88), and Mask R-CNN (0.90); for bacterial black spot, YOLOv11 (0.31), SSD (0.69), RetinaNet (0.85),

Faster R-CNN (0.84), and Mask R-CNN (0.82); for others, YOLOv11 (0.51), SSD (0.76), Faster R-CNN (0.84), and both RetinaNet and Mask R-CNN (0.92). Fruitfly detection attained lowest performance, with YOLOv11 (0.06), SSD (0.14), RetinaNet (0.39), Mask R-CNN (0.46), and Faster R-CNN (0.59). Generally, RetinaNet and Mask R-CNN showed the most consistent performance.

C. Computational complexity analysis of detection models

Computational complexity varied across models: SSD (24.1M parameters, 30 GFLOPs, 9.0 ms inference time) and YOLOv11 (20.0M, 35 GFLOPs, 9.5 ms) had the lowest cost, making them most suitable for real-time use, with RetinaNet also attaining moderate complexity (36.4M, 130 GFLOPs, 23 ms). In contrast, two-stage models Faster R-CNN (43.3M, 280 GFLOPs, 35 ms) and Mask R-CNN (45.9M, 300 GFLOPs, 57 ms) had the highest computational demands, reflecting the trade-off between accuracy and efficiency.

TABLE I
FAITHFULNESS SCORES ACROSS MODELS AND XAI TECHNIQUES

Model	Grad-CAM	Grad-CAM++	LRP	LIME	SHAP
YOLOv11	0.58	0.30	0.76	0.53	0.56
SSD	0.58	0.30	0.80	0.53	0.73
RetinaNet	0.33	0.33	0.33	0.33	0.33
Faster R-CNN	0.41	0.41	0.51	0.41	0.46
Mask R-CNN	0.34	0.38	0.52	0.38	0.41

TABLE II
ROBUSTNESS SCORES ACROSS MODELS AND XAI TECHNIQUES

Model	Grad-CAM	Grad-CAM++	LRP	LIME	SHAP
YOLOv11	0.00	0.00	0.22	0.00	0.00
SSD	0.62	0.86	0.59	0.00	0.00
RetinaNet	0.50	0.50	0.43	0.00	0.00
Faster R-CNN	0.41	0.81	0.43	0.00	0.00
Mask R-CNN	0.43	0.82	0.31	0.00	0.00

TABLE III
LOCALIZATION SCORES ACROSS MODELS AND XAI TECHNIQUES

Model	Grad-CAM	Grad-CAM++	LRP	LIME	SHAP
YOLOv11	0.00	0.00	0.60	0.11	0.16
SSD	0.13	0.05	0.49	0.15	0.43
RetinaNet	0.00	0.00	0.00	0.01	0.00
Faster R-CNN	0.12	0.07	0.30	0.00	0.18
Mask R-CNN	0.16	0.06	0.33	0.00	0.52

D. Visualization models

To improve model interpretability, XAI techniques were integrated with the deep learning models to highlight influential image regions and quantify feature contributions to model predictions giving comprehensive visual insights into the decision-making process of the detection models, a sample is illustrated in Figure 1. XAI visualizations revealed how model predictions align with ground truth, highlighting both agreement and mismatch cases. Attribution maps highlighted

key lesion regions that most influenced model decision in doing predictions, while attention heatmaps were used to show how much of the model's focus lies within ground truth bounding boxes, indicating attention to relevant areas. However, this focus is often limited, and thus this may lead to missed detections when multiple damage instances are present on the fruit.

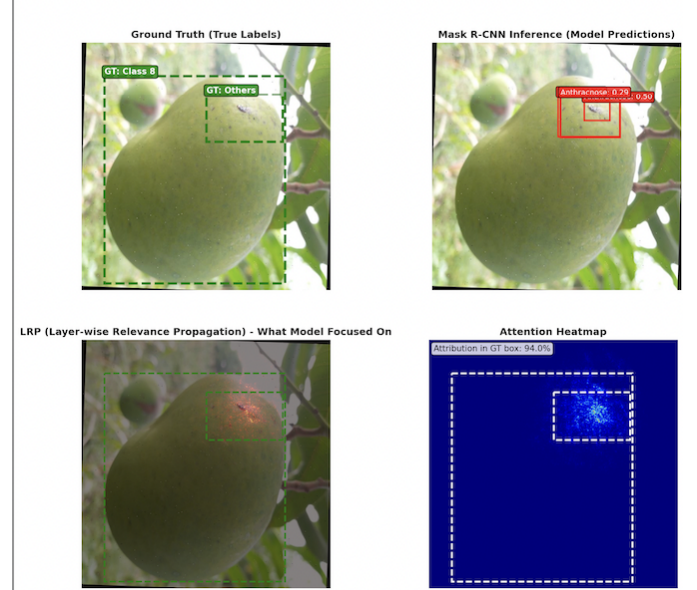


Fig. 1. Visualization of Mask R-CNN Predictions and LRP-Based Attribution for Mango Damage Detection

E. Quantitative Evaluation of the XAI Techniques

The performance of XAI techniques was evaluated using faithfulness, robustness, complexity, localization, and randomization, as reflected in Tables I, II, and III with variations in results attributed to differences in their underlying mechanisms and sensitivity to feature granularity. Grad-CAM and Grad-CAM++ use gradient-weighted feature maps from the final convolutional layer to highlight key regions influencing model decisions, making them robust to diverse damage textures and lesion patterns [21] and results attained as seen in Table II are in agreement with this. Results show that LRP and SHAP achieved higher faithfulness and localization scores, indicating accurate and meaningful feature contributions aligned with ground truth. In contrast, Grad-CAM and Grad-CAM++ demonstrate poor performance across most metrics, particularly for YOLOv11, where zero values were observed in all measures except for faithfulness. This suggests an inability to generate well-localized explanations, where the highlighted regions may influence model predictions but do not correspond to relevant features. Additionally, Grad-CAM++ shows higher robustness in some models, indicating more stable explanations under small input perturbations. The results show that LIME generates highly complex explanations, capturing excessive detail and thus making interpretation difficult. Overall, the results attained show that XAI methods have varying

strengths and limitations, highlighting the need for multi-metric evaluation criteria of explanation quality rather than relying on a single metric.

V. DISCUSSION OF STUDY FINDINGS

The results attained indicate that the evaluated models effectively detect, localize, and identify causes of mango damage, with RetinaNet attaining the best precision, recall, and F1-score. The variations in performance across models are attributed to their architectural design and learning strategies. The DL models that handle class imbalance achieved higher detection accuracy. Across all models, there were consistently high IoU values attained, further indicating that all models employed in this study are capable of accurate localisation. Per-class accuracy was strong across most categories, except for fruitfly, which remained challenging possibly due to its small size and subtle features. The integration of XAI techniques to detection models reveals that models generally focus on relevant lesion areas of the fruit, although attention is at times limited, which brings about incomplete detections. The quantitative XAI results show that model architecture affects explainability, with two-stage detectors providing more localized explanations, while Grad-CAM based methods performs poorly on YOLOv11. Despite poor localization, faithfulness attained indicates that explanations may influence model decision-making on predictions without accurately reflecting relevant regions, while high complexity limits interpretability.

VI. CONCLUSION AND FUTURE WORK

This study used deep learning to detect mango damage and its causes, showing effective results, though improvements are still needed. XAI integration provided interpretable insights into model predictions, highlighting their strengths and limitations of methods used, and demonstrating the potential of combining object detection models with explainability for practical agricultural applications. Given that image fusion-based models outperform single-image approaches [22], future work can explore this option to improve mango damage detection and cause identification. Future work will further explore integrating attention mechanisms and transformer-based models with XAI models, as prior studies have demonstrated good performance with such use of such methods [6].

REFERENCES

- [1] Arshleen Kaur, Vinay Kukreja, Deepak Singh Rana, Ashish Garg, and Rishabh Sharma. Deep insights into mango scab: An ai-powered approach for detection and classification. In *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–5, 2024.
- [2] Sandeep Kumar, Bhupesh Gupta, Mohit Chhabra, and Lucy Garg. Comparative evaluation of mango fruit diseases classification using machine learning. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)*, pages 581–585, 2024.
- [3] Archana Ganesh Said and Bharti Joshi. Advanced multimodal thermal imaging for high-precision fruit disease segmentation and classification. *Journal of Autonomous Intelligence*, 7, 2024.
- [4] Xinxing Zhang, Yi Xun, and Yaohui Chen. Automated identification of citrus diseases in orchards using deep learning. *Biosystems Engineering*, 223:249–258, 2022.
- [5] M. Aldiki Febriantono. Xai-driven apple disease identification using efficientnet and grad-cam. In *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, pages 1–6, 2025.
- [6] Nagineni Venkata Sireesha, Gillala Rekha, and Venkataramana Gun-treddi. Interpretable citrus disease classification through gradcam++ lime and vision transformer-based deep learning. *Discover Applied Sciences*, 2026.
- [7] Sahitya Mondal, Tapashri Sur, Diganta Sengupta, and Chitrita Chaudhuri. Mobilenet based fruit classification using unet generated segmented images: Grad-cam visualization. In *International Conference on Applied Algorithms*, pages 237–249. Springer, 2025.
- [8] Priya Singh, Ajay Rana, Tanupriya Choudhury, and Nandita Tripathi. Deep feature fusion and explainable ai for vision-based mango defect classification. *Intelligent Data Analysis*, page 1088467X261433368, 2025.
- [9] Yunong Tian, Guodong Yang, Zhe Wang, En Li, and Zize Liang. Detection of apple lesions in orchards based on deep learning methods of cyclegan and yolov3-dense. *Journal of Sensors*, 2019(1):7630926, 2019.
- [10] Ramazan Hadipour-Rokni, Ezzatollah Askari Asli-Ardeh, Ahmad Jahanbakhshi, Sajad Sabzi, et al. Intelligent detection of citrus fruit pests using machine vision system and convolutional neural network through transfer learning technique. *Computers in Biology and Medicine*, 155:106611, 2023.
- [11] ASM Farhan Al Haque, Rubaiya Hafiz, Md Azizul Hakim, and GM Rasiqul Islam. A computer vision system for guava disease detection and recommend curative solution using deep learning approach. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE, 2019.
- [12] Ibrahim Alrashdi, Marwa Sharawi, Ahmed M Ali, Karam M Sallam, Bilal Arain, and Mohamed Abdel-Basset. Utilizing deep learning models for early detection and classification of fruit diseases: towards sustainable agriculture and enhanced food quality. *Scientific Reports*, 2026.
- [13] Na Yao, Fuchuan Ni, Minghao Wu, Haiyan Wang, Guoliang Li, and Wing-Kin Sung. Deep learning-based segmentation of peach diseases using convolutional neural network. *Frontiers in Plant Science*, 13:876357, 2022.
- [14] Nidhal Jegham, Chan Young Koh, Marwan Abdelatti, and Abdeltawab Hendawi. Yolo evolution: A comprehensive benchmark and architectural review of yolov12, yolo11, and their previous versions. *arXiv preprint arXiv:2411.00201*, 2024.
- [15] Karel Horak and Robert Sablatnig. Deep learning concepts and datasets for image recognition: overview 2019. In *Eleventh international conference on digital image processing (ICDIP 2019)*, volume 11179, pages 484–491. SPIE, 2019.
- [16] Sana Z. Khan, Salam Dhou, and A. R. Al-Ali. Machine learning based palm farming: Harvesting and disease identification. *IEEE Access*, 12:157854–157871, 2024.
- [17] Vaibhav Dhore, Achintya Bhat, Viraj Nerlekar, Kashyap Chavhan, and Aniket Umare. Enhancing explainable ai: a hybrid approach combining gradcam and lrp for cnn interpretability. *arXiv preprint arXiv:2405.12175*, 2024.
- [18] Yonasi Safari, Joyce Nakatumba-Nabende, Rose Nakasi, and Rose Nakibuule. A review on automated detection and assessment of fruit damage using machine learning. *IEEE access*, 12:21358–21381, 2024.
- [19] Philine Lou Bommer, Marlene Kretschmer, Anna Hedström, Dilyara Bareeva, and Marina M-C Höhne. Finding the right xai method—a guide for the evaluation and ranking of explainable ai methods in climate science. *Artificial Intelligence for the Earth Systems*, 3(3):e230074, 2024.
- [20] Jonas Klotz, Tom Burgert, and Begüm Demir. On the effectiveness of methods and metrics for explainable ai in remote sensing image scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:27764–27780, 2025.
- [21] Md Naimul Islam Nuhash, Md Sohag, Shahriar Sultan Ramit, and Raja Tariqul Hasan Tusher. A deep ensemble learning and explainable ai framework for accurate bottle gourd disease diagnosis and deployment. *Smart Agricultural Technology*, page 101541, 2025.
- [22] Luis Chuquimarca, Boris Vintimilla, and Sergio Velastin. Classifying healthy and defective fruits with a multi-input architecture and cnn models. In *2024 14th International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–7. IEEE, 2024.