ISBN: 978-628-96613-1-6. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2025.1.1.1781

Dataset validation for Disease Detection in Tomato Plants

Jose Luis Ordoñez-Avila, Phd en Dirección Empresaria[1], William Fajardo, Ingeniero en Mecatrónica[1], Mauro Escobar, Ingeniero en Mecatrónica[1], Douglas Aguilar, Master en Robótica[2], David C. Balderas, Phd en Inteligencia Artificial[3]
[1]Facultad de Ingeniería, Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras,
[2]Facultad de Ingeniería, Universidad Evangélica de El Salvador (UEES), San Salvador, El Salvador,
[3]Institute of Advanced Materials for Sustainable Manufacturing, Tecnologico de Monterrey, Anillo Perif. 6666, México City 14380, México
jlordonez@unitec.edu, williamisaac@unitec.edu, mauro.escobar@unitec.edu, douglas.aguilar@uees.edu.sv, dc.balderassilva@tec.mx

Abstract– Tomato cultivation is a vital agricultural activity worldwide, contributing significantly to global food production. However, tomato crops are highly susceptible to various diseases, including mold, bacterial spot, and early blight, which can severely impact fruit quality and yield. These diseases, if not detected and managed promptly, lead to increased production costs and decreased efficiency. This research aims to address these challenges by developing and implementing an early disease detection dataset using Convolutional Neural Networks (CNNs). The system was trained with 4,083 images of tomato plants, allowing the CNN model to accurately identify specific diseases in both early and advanced stages. The model achieved a mean Average Precision (mAP) of 86.1%, a precision of 88.2%, and a recall of 82.6%, indicating its effectiveness of the dataset. This dataset can be used to develop different applications for managing tomatoes farm.

Keywords-- List at most 5 key index terms here.

I. INTRODUCTION

Tomatoes are widely cultivated due to their nutritional value and profitability for farmers, contributing to global nutrition [1]. Their economic and nutritional importance makes them one of the most widely grown crops globally [2]. The most common diseases in this plant are the mold and bacterial spots which spread rapidly and can seriously affect the crop. The presence of these diseases not only reduces crop productivity, but also significantly increases production costs, negatively affecting growers' profitability.

Crop yield, essential for agricultural sustainability, can be affected by plant diseases [3]. Recognizing these diseases is important to take timely action and improving crop growth [4]. Identifying and controlling diseases without the need for specialists enhances both the quality and quantity of tomato production, economically benefiting farmers [5]. Tomatoes, which are highly nutritious, have a significant impact on the agricultural economy [6]. In rural areas, millions depend on agriculture as their primary source of income [7]. Moreover, it is crucial to continuously monitor plant health to anticipate its impact in the field. Agriculture is one of the main goals to global sustainability and subsistence [8].

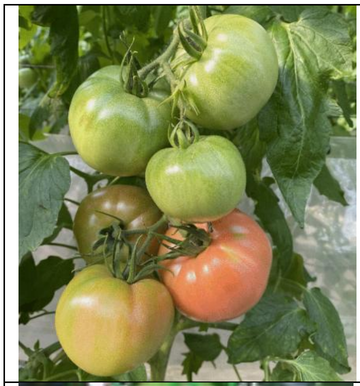
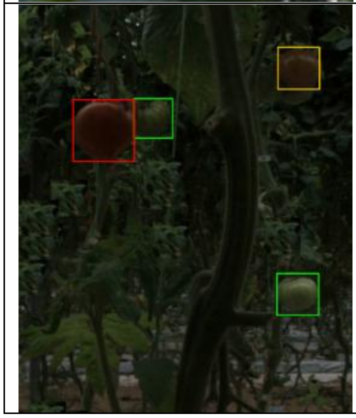
Image analysis is used to identify species, classify fruits, and diagnose plant diseases [9-11]. Early detection of diseases in plants can decrease dependence on dangerous chemicals for their growth and protection [12]. Agriculture relies on the identification of plant diseases to prevent their spread and ensure a high-quality harvest. Farmers may overapply pesticides when they fail to properly identify diseases, harming the plants and reducing productivity. Early disease detection reduces the need for hazardous chemicals and allows for a more effective response to protect plants and the environment.

Farmers can use this method to determine when tomatoes are at their ideal ripeness and if it is the right time for harvest [13]. In a conventional image classification network, feature extraction is performed uniformly across the entire image, regardless of the proportion of the key discriminant region [14]. Traditional disease detection methods involve manual inspection of diseased leaves through visual cues or chemical analysis of affected areas, which can lead to low detection efficiency and limited reliability due to human error [15]. Agricultural automation using technologies like image classification networks can reduce dependence on manual inspection for detecting plant diseases. Faster and more accurate detection, along with the ability to make data-driven decisions in real time, are some of the additional benefits gained from integrating these technologies with technology platforms and cloud computing.

Efficient detection and management of crop diseases can increase yield and quality and minimize resource waste [16]. Otherwise, a massive outbreak of diseases can devastate previously established crops, causing devastating and irreparable losses [17]. Table I shows examples of different datasets with their number of images and examples of tomato detection.

Table 1. Data sets examples.

Sample	Images	Reference

	804	[18]
	449	[19]
	277	[20]

Agricultural production is negatively affected by diseases that attack plants. Organic agents and other pathogens can cause them. To understand crop development and take swift action, it is essential to identify these diseases on leaves. Crop yield and quality can be improved, resource waste reduced, and agroecosystems preserved through effective disease detection and management. This project focuses on the validation of the number of images in the dataset and the dataset itself can be useful for tomato farm research, detecting green tomatoes and plant disease. The main objective of this work is to validate the number of images in this dataset and the dataset itself can be useful for tomato farm research.

This paper is divided into several sections, starting with the methods section which explains how the dataset was used for each of the three experiments. The results section presents figures that validate the training of the dataset and its respective metrics. Finally, we conclude on the number of images and the usability of this dataset.

II. Methods

For the development of this study, a dataset composed of images of tomato plants in different health conditions was used. Three experiments were defined to evaluate the detection of healthy tomatoes and tomato plants with mold and bacterial spots as shown in Table II.

Table II. Objects and images experiments.

Experiment	Class	Object	Images
1	1	Healthy Tomatoes	2384
2	1 2	Healthy Tomatoes Mold	3444
3	1 2 3	Healthy Tomatoes Mold Bacterial Spot	4083

All images in the dataset were taken in an open field, which introduces a higher level of noise compared to images captured in controlled laboratory environments. For data collection, the team traveled to a tomato farm in the town of El Rosario, Comayagua, Honduras, with the specific objective of capturing photographs under real growing conditions. This approach allowed generating a dataset more representative of practical scenarios, facilitating the use of convolutional neural networks (CNN) for any user without requiring extensive training.

Fig. 1 Geographical Location of Tomato Farm in El Rosario Village, Comayagua, coordinates 14.62555, -87.76977.

The dataset was processed and annotated using RoboFlow, a platform that facilitates the management of object annotations. Figure 2 shows an example of class 1, class 2 and class 3 annotations. Proper labeling and image resolution are factors for ensuring effective model training. Subsequently, the dataset was divided into three parts:

- Training (70%): this subset represents the bulk of the data, as the model needs enough information to learn relevant and generalizable patterns in detecting healthy and diseased tomatoes.

- Testing (20%): Used to evaluate the performance of the model after training, providing an estimate of its generalizability to previously unseen data.
- Validation (10%): Intended for model optimization during training, allowing adjustments to the hyperparameters and avoiding problems such as overfitting.

Experiment 1 focused on the detection of healthy tomatoes, using 2,384 images. Experiments 2 and 3 analyzed the detection of tomatoes with mold and bacterial spots, using 3,444 and 4,083 images, respectively. The difference in the number of images between the latter allowed evaluating the impact of the dataset size on the model performance.

For the evaluation of the dataset, the distribution of the annotations was analyzed, and an increase of images was performed in each experiment to appreciate relevant changes in the indicators. The most relevant indicators were precision, recall, mean average precision (mAP), box loss, object loss and class loss. These experiments will allow us to understand how the increase of classes and images affect the classification efficiency of the network and to conclude if the dataset has enough images and annotations to be used in artificial intelligence applications.

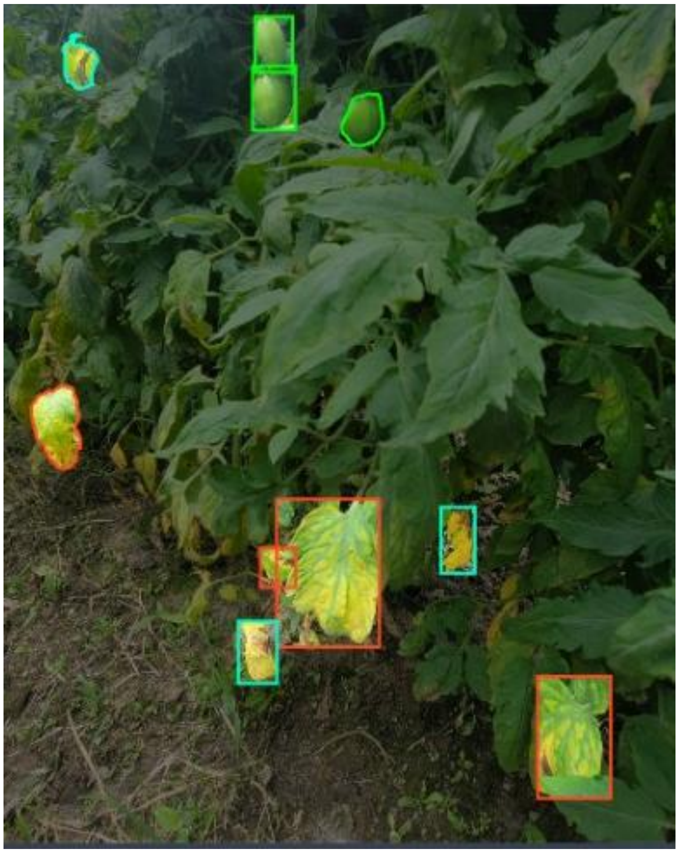


Fig. 2 Anotation example in roboflow: green box) class 1; red box) class 2; magenta box) class 3.

IV. Analyss and Results

In this section the results will be analyzed in detail for each process performed in the research, showing the variations between the different increments implemented. This will allow us to evaluate how image addition, model parameter adjustments and labeling improvements influenced the performance of the Convolutional Neural Network (CNN). In addition, key metrics such as mAP, precision and recall will be examined to understand how each adjustment optimized disease detection in tomatoes.

A. *Experiment 1: Green Tomatoes*

The results obtained after training in Roboflow provide a detailed view of the model's performance in detecting green tomatoes. Throughout this process, it was possible to evaluate the ability of the Convolutional Neural Network to adapt to different field scenarios, addressing challenges such as variations in lighting, shooting angle, and the diversity of visual characteristics of the tomatoes. Figure 3 shows the frequency of annotations per image with a median resolution of 1480x1920 pixels. The highest frequency ranges from 2 to 14 annotations per image of green tomatoes and the average per image was 14.9 annotations. This frequency exceeds 1000 images and therefore makes up approximately 50% of the images used for this test. The box loss 0.912, class loss 0.511, and object loss 0.966 in this scenario were trending less than 1 (fig. 4) which is beneficial to the network. This shows that with 2384 images out of 4083, the neural network was able to detect with high accuracy. The final indicators of the neural network were 93.6% mAP, 93.1% precision and 87.1% recall.

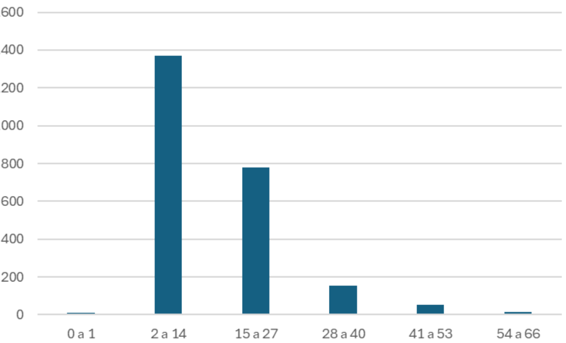


Fig. 3 Number of images, and label frequency for experiment 1, total annotations 35630.

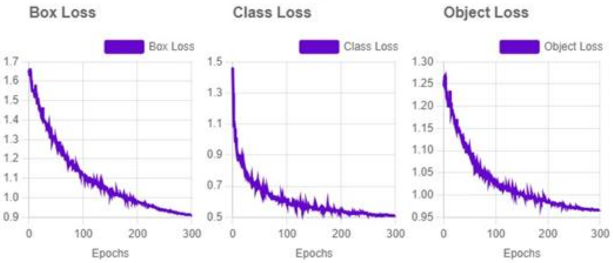


Fig. 4 Box Loss, Class Loss and Object Loss for experiment 1.

B. Experiment 2: Mold and Bacterial Spot Disease

During experiment 2, additional images were added to the scenario 1 dataset. In this phase, 3,444 accurately labeled images were used to identify mold and bacterial spot-on tomato leaves. Figure 5 shows the frequency of annotations per image; the median resolution was 1480x1920 pixels. The highest frequency ranges from 2 to 16 annotations per image of disease detections. This number of images represents more than 2/3 of the whole data set.

The mAP for experiment 2 reached a value of 88.4%, reflecting a low decrease in comparison to experiment 1. This low decrease. This drop in the result is due to the difficulty in detecting diseases, especially when there are two different diseases. This increases the confusion of the network to perform the classification, however the result is acceptable and will allow it to be applied with other images. Figure 6 presents the losses during the training process of experiment 2. The Box Loss, which measures the error in predicting object boundaries, was 0.832. The Class Loss, which indicates the error in classifying detected objects, reached 0.569, while the Object Loss, which measures the certainty of an object's existence in a region, was 0.954.

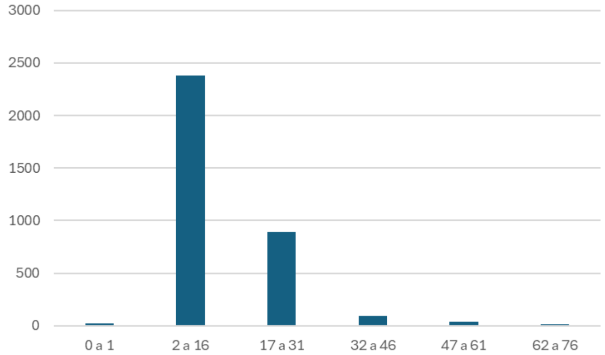


Fig. 5 Number of images, and label frequency for experiment 2, total annotations 36057 for class 1 and 12215 for class 2.

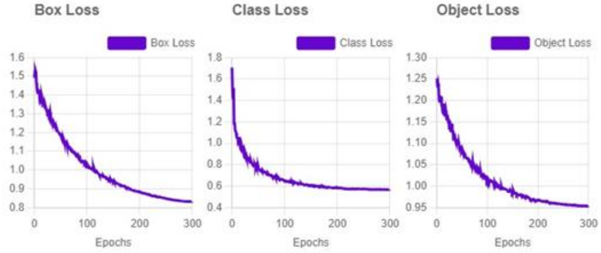


Fig. 6 Box Loss, Class Loss and Object Loss for experiment 2.

C. Experiment 3: Mold and Bacterial Spot Disease

In this experiment, we evaluated how many images and annotations were necessary for effective disease detection. During this stage, additional images were added to the scenario 2 dataset. In this phase, 4083 accurately labeled images were used to improve the model's ability to identify mold and bacterial spot-on tomato leaves. Figure 7 shows the frequency of annotations per image; the median resolution was 1480x1920 pixels. The highest frequency ranges from 2 to 18 annotations per image of disease detections.

The mAP (Mean Average Precision) improved over the epochs of training during experiment 3, reaching a final value of 86.1%, which reflects the model's ability to accurately detect diseases on tomato leaves. The losses associated with the model during the training of this experiment are observed. The Box Loss is at 0.648, Class Loss at 0.516, and Object Loss at 0.919, reflecting continuous improvement in disease prediction and the correct classification of detected objects.

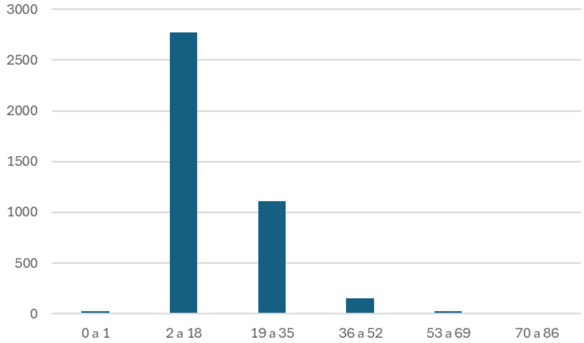


Fig. 7 Number of images, and label frequency for experiment 3, total annotations 44201 for class 1, 13959 for class 2, and 7478 for class 3.

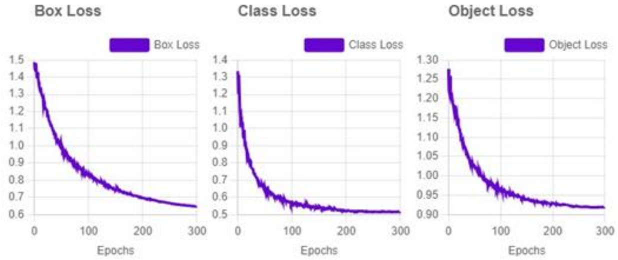


Fig. 8 Box Loss, Class Loss and Object Loss for experiment 3.

D. Experiments summary

The results obtained after the three experiments are shown in Table III, including annotations. Class 2 and 3 are underrated in comparison with class 1, that is why the best model was ensured in experiment 1. The dataset has allowed the trained models to perform acceptably in the detection of healthy tomatoes, with a mAP of 93.6%, indicating high precision and recall in this category. In the detection of tomatoes with mold and bacterial spots, the performance is slightly lower (mAP of 88.4% and 86.1% in experiments 2 and 3, respectively), with a slight decrease in precision and recall. The difference between these two experiments is not significant, suggesting that adding more images to the training did not substantially improve performance, possibly due to the complexity of visual variability in the affected tomatoes (different types of spots, colors and textures) or the presence of redundant data.

This dataset without the use of augmentations and without parameter variations that could substantially improve the results presents acceptable indicators, so it is useful for other researchers to make use of the dataset. Figure 9 is an example of the detections made with validation images, which were not used for training, in these can be seen the detections and non-detections of some elements. Although some objects were not detected correctly, the dataset can be improved by adjusting the hyperparameters of the model.

The three-dimensional identification of tomatoes are essential for robotic harvesting, and its effectiveness depends on the speed and accuracy of the vision system [21]. A lightweight CNN model is presented to obtain high-level hidden features [22]. This approach uses machine learning and a lightweight CNN model to enhance visual collection and plant health monitoring. Obtaining precise solutions improves the efficiency of harvesting robots and enhances crop management by detecting diseases and deficiencies, facilitating its use in the field, especially for low-performance mobile devices [23].

Table III. Results.

Experiment	Object	Images	Annotations	mAP	Precision	Recall
1	Healthy Tomatoes	2384	35630	93.6%	93.1%	87.1%
2	Healthy Tomatoes and Mold	3444	48272	88.4%	89.9%	83.0%
3	Healthy Tomatoes, Mold and Bacterial Spot	4083	65638	86.1%	88.2%	82.6%

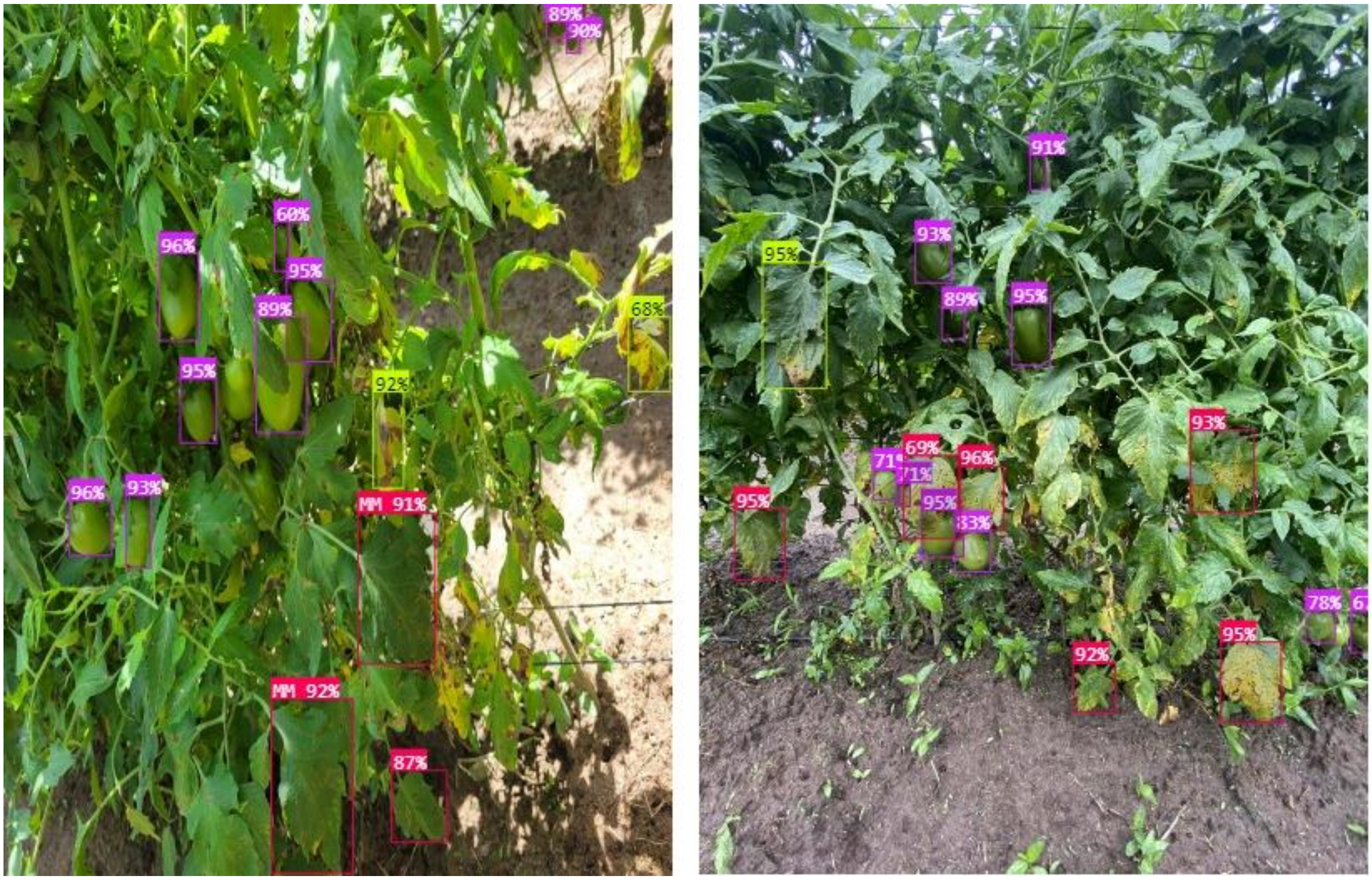


Fig. 9 Example of images in the proposed dataset (download here).

CONCLUSIONS

After performing three experiments for the detection of healthy tomatoes and two tomato diseases, acceptable results were obtained. The model has the best performance in detecting healthy tomatoes, with a mAP of 93.6%, indicating high precision and recall in this category. For tomatoes with mold and bacterial spots, the performance is lower (mAP of 88.4% and 86.1% in experiments 2 and 3, respectively), with a slight decrease in precision and recall. Although the number of images and annotations increased between experiments 2 and 3, the improvement in the model was not significant, suggesting that the complexity of the problem lies not only in the amount of data, but in the visual variability of the lesions and the quality of the labeling.

The dataset used with a considerable volume of images and annotations, could be useful for other research related to crop disease detection using computer vision. Its application could be extended to the identification of other diseases in tomatoes or even other agricultural products, by training new models with transfer learning techniques. In addition, the analysis of disease evolution at different stages of tomato growth could be useful in studies on the impact of environmental conditions or phytosanitary control strategies.

Acknowledgment

We would like to express our sincere gratitude to the tomato farm located in Comayagua for opening their doors to us and allowing us to conduct the photo-taking process on their premises.

References

[1] M. Umar, S. Altaf, S. Ahmad, H. Mahmoud, A. S. N. Mohamed, y R. Ayub, «Precision Agriculture Through Deep Learning: Tomato Plant Multiple Diseases Recognition With CNN and Improved YOLOv7», IEEE Access, vol. 12, pp. 49167-49183, 2024, doi: 10.1109/ACCESS.2024.3383154.

[2] E. Özbılge, M. K. Ulukök, Ö. Toygar, y E. Ozbılge, «Tomato Disease Recognition Using a Compact Convolutional Neural Network», IEEE Access, vol. 10, pp. 77213-77224, 2022, doi: 10.1109/ACCESS.2022.3192428.

[3] H. D. Gadade y D. K. Kirange, «Tomato Leaf Disease Diagnosis and Severity Measurement», 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), pp. 318-323, jul. 2020, doi: 10.1109/WorldS450073.2020.9210294.

[4] C. Zhou, S. Zhou, J. Xing, y J. Song, «Tomato Leaf Disease Identification by Restructured Deep Residual Dense Network», IEEE Access, vol. 9, pp. 28822-28831, 2021, doi: 10.1109/ACCESS.2021.3058947.

[5] H. I. Peyal et al., «Plant Disease Classifier: Detection of Dual-Crop Diseases Using Lightweight 2D CNN Architecture», IEEE Access, vol. 11, pp. 110627-110643, 2023, doi: 10.1109/ACCESS.2023.3320686.

[6] R. Satya Rajendra Singh y R. K. Sanodiya, «Zero-Shot Transfer Learning Framework for Plant Leaf Disease Classification», IEEE Access, vol. 11, pp. 143861-143880, 2023, doi: 10.1109/ACCESS.2023.3343759.

[7] Q. Wu, Y. Chen, y J. Meng, «DCGAN-Based Data Augmentation for Tomato Leaf Disease Identification», IEEE Access, vol. 8, pp. 98716-98728, 2020, doi: 10.1109/ACCESS.2020.2997001.

[8] K. Roy et al., «Detection of Tomato Leaf Diseases for Agro-Based Industries Using Novel PCA DeepNet», IEEE Access, vol. 11, pp. 14983-15001, 2023, doi: 10.1109/ACCESS.2023.3244499.

[9] M. S. Fuentes, N. A. L. Zelaya, y J. L. O. Avila, «Coffee Fruit Recognition Using Artificial Vision and neural NETWORKS», en 2020 5th International Conference on Control and Robotics Engineering (ICCRE), abr. 2020, pp. 224-228. doi: 10.1109/ICCRE49379.2020.9096441.

[10] R. Espinal-Lanza, M. E. Perdomo, J. Sanchez-Palma, y J. L. Ordoñez-Avila, «Comparison of Deep Learning Technologies Applied to the Recognition of Defects in Cocoa Beans», en 2023 IEEE 41st Central America and Panama Convention (CONCAPAN XLI), nov. 2023, pp. 1-6. doi: 10.1109/CONCAPANXLI59599.2023.10517568.

[11] E. M. T. Caballero y A. M. R. Duke, «Implementation of Artificial Neural Networks Using NVIDIA Digits and OpenCV for Coffee Rust Detection», en 2020 5th International Conference on Control and Robotics Engineering (ICCRE), abr. 2020, pp. 246-251. doi: 10.1109/ICCRE49379.2020.9096435.

[12] W. Shafik, A. Tufail, A. Namoun, L. C. De Silva, y R. A. A. H. M. Apong, «A Systematic Literature Review on Plant Disease Detection: Motivations, Classification Techniques, Datasets, Challenges, and Future Trends», IEEE Access, vol. 11, pp. 59174-59203, 2023, doi: 10.1109/ACCESS.2023.3284760.

[13] Y.-P. Huang, T.-H. Wang, y H. Basanta, «Using Fuzzy Mask R-CNN Model to Automatically Identify Tomato Ripeness», IEEE Access, vol. 8, pp. 207672-207682, 2020, doi: 10.1109/ACCESS.2020.3038184.

[14] Y. Wu, X. Feng, y G. Chen, «Plant Leaf Diseases Fine-Grained Categorization Using Convolutional Neural Networks», IEEE Access, vol. 10, pp. 41087-41096, 2022, doi: 10.1109/ACCESS.2022.3167513.

[15] S. Ahmed, Md. B. Hasan, T. Ahmed, Md. R. K. Sony, y Md. H. Kabir, «Less is More: Lighter and Faster Deep Neural Architecture for Tomato Leaf Disease Classification», IEEE Access, vol. 10, pp. 68868-68884, 2022, doi: 10.1109/ACCESS.2022.3187203.

[16] J. Feng, W. E. Ong, W. C. Teh, y R. Zhang, «Enhanced Crop Disease Detection With EfficientNet Convolutional Group-Wise Transformer», IEEE Access, vol. 12, pp. 44147-44162, 2024, doi: 10.1109/ACCESS.2024.3379303.

[17] G. Yang, G. Chen, Y. He, Z. Yan, Y. Guo, y J. Ding, «Self-Supervised Collaborative Multi-Network for Fine-Grained Visual Categorization of Tomato Diseases», IEEE Access, vol. 8, pp. 211912-211923, 2020, doi: 10.1109/ACCESS.2020.3039345.

[18] Roman E., Dewen X., Fujihara H. (2020). LabotoTomato: Laboro Tomato: Instance segmentation dataset. Available from https://datasetninja.com/ag-rob-tomato

[19] Roman E., Dewen X., Fujihara H. (2020). LabotoTomato: AgRobTomato Dataset: Greenhouse tomatoes with different ripeness stages. Zenodo. Available from https://doi.org/10.5281/zenodo.5596799

[20] Tsironis V., Bourou S., Stentoumis C. (2020). tomatOD: Evaluation of object detection algorithms on a new real-world tomato dataset. In ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Available from https://github.com/up2metric/tomatOD

[21] C. Song et al., «TDPPL-Net: A Lightweight Real-Time Tomato Detection and Picking Point Localization Model for Harvesting Robots», IEEE Access, vol. 11, pp. 37650-37664, 2023, doi: 10.1109/ACCESS.2023.3260222.

[22] K. M. Hosny, W. M. El-Hady, F. M. Samy, E. Vrochidou, y G. A. Papakostas, Multi-Class Classification of Plant Leaf Diseases Using Feature Fusion of Deep Convolutional Neural Network and Local Binary Pattern», IEEE Access, vol. 11, pp. 62307-62317, 2023, doi: 10.1109/ACCESS.2023.3286730.

[23] J. L. Ordonez Avila y J. Fernandez-Olivares, «A Mobile Robot with Deep Learning for Monitoring Coffee Farms», en Proceedings of the 22nd LACCEI International Multi-Conference for Engineering, Education and Technology (LACCEI 2024): "Sustainable Engineering for a Diverse, Equitable, and Inclusive Future at the Service of Education, Research, and Industry for a Society 5.0.", Latin American and Caribbean Consortium of Engineering Institutions, 2024. doi: 10.18687/LACCEI2024.1.1.1924.