British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 1, 2026 (pp. 137-146)
www.abjournals.org

CocoaDetectDB: A TinyML-ORIENTED IMAGE DATASET FOR COCOA PLANT DISEASE DETECTION

Bassey Isaac Rajuno (Ph.D.)[1] and Ekoro Ekoro Igo (Ph.D.)[2].

[1]Department of Computer Science, College of Health Technology, Calabar.
Email: isaacrajuno@gmail.com; Tel.: +2348037937122

[2]Department of Computer Science, University of Education and Entrepreneurship, Akamkpa.
Email: eekoro@gmail.com, ekoro.ekoro@crs-coeakamkpa.edu.ng; Tel.: +2348039516783

Cite this article:

Bassey, I. R., Ekoro, E. I. (2026), CocoaDetectDB: A TinyML-Oriented Image Dataset for Cocoa Plant Disease Detection. British Journal of Computer, Networking and Information Technology 9(1), 137-146. DOI: 10.52589/BJCNIT-NP2MMBZN

Manuscript History

Received: 18 Jan 2026

Accepted: 20 Feb 2026

Published: 16 Apr 2026

ABSTRACT: *The application of computer vision in precision agriculture has demonstrated considerable promise in automated plant disease detection. However, the effectiveness of such approaches is strongly dependent on the availability of high-quality, domain-specific datasets, particularly for deployment on resource-constrained edge devices. This paper introduces CocoaDetectDB, a publicly available image dataset developed for the detection of cocoa plant diseases under Tiny Machine Learning (TinyML) constraints. The dataset comprises images of healthy cocoa pods and three major cocoa diseases—Cocoa Black Pod Disease (CBD), Cocoa Swollen Shoot Virus Disease (CSSVD), and Frosty Pod Rot (FPR)—captured under real-world field conditions and supplemented with openly accessible public data. Images were curated, cleaned, and resized to a uniform resolution of 112 × 112 pixels to support low-memory and low-power inference. To validate the suitability of the dataset for automated disease classification, baseline experiments were conducted using MobileNetV2 and a lightweight quantized TensorFlow Lite model. Experimental results demonstrate classification accuracies of 99.13% and 93.75%, respectively, indicating that CocoaDetectDB contains sufficiently discriminative features for both conventional lightweight models and TinyML deployment. The dataset is intended to support future research in cocoa disease detection, edge AI, and resource-efficient agricultural monitoring systems.*

KEYWORDS: Cocoa Disease Detection, Agricultural Datasets, TinyML, Computer Vision, Edge AI, TensorFlow Lite.

Article DOI: 10.52589/BJCNIT-NP2MMBZN
DOI URL: https://doi.org/10.52589/BJCNIT-NP2MMBZN

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 1, 2026 (pp. 137-146)
www.abjournals.org

INTRODUCTION

In recent years, artificial intelligence (AI) techniques have been increasingly adopted in precision agriculture to improve crop monitoring, disease diagnosis, and yield optimization. Among these techniques, deep learning—particularly convolutional neural networks (CNNs)—has played a dominant role in image-based plant disease detection, achieving high levels of accuracy across various benchmark datasets (Liu & Wang, 2021; Ren, Kim, & Jeong, 2020). CNN-based models have demonstrated performance comparable to or exceeding human-level accuracy on several large-scale image classification benchmarks (He, Zhang, Ren, & Sun, 2015), reinforcing their applicability to agricultural vision tasks.

Despite these advances, the success of deep learning models is heavily dependent on the availability of large, annotated, and domain-relevant datasets. In agricultural contexts, such datasets remain limited, especially for region-specific crops and diseases (Rodriguez, Alfaro, Paredes, Esenarro, & Hilario, 2021; Syamsuri & Kusuma, 2019). Zheng et al. (2019a) emphasized the urgent need for annotated crop vision datasets tailored to specific agricultural domains in order to enable robust and generalizable model development.

Cocoa (Theobroma cacao) is a crop of significant economic importance, particularly in West Africa, which accounts for a substantial proportion of global cocoa production. Cocoa plants are susceptible to several devastating diseases, including Cocoa Black Pod Disease (CBD), Cocoa Swollen Shoot Virus Disease (CSSVD), and Frosty Pod Rot (FPR). Early and accurate detection of these diseases is critical for effective disease management and yield preservation. However, publicly available image datasets dedicated to cocoa plant diseases remain scarce, and none explicitly address the constraints of TinyML-oriented deployment.

In response to this gap, this paper introduces CocoaDetectDB, a domain-specific image dataset designed for cocoa plant disease detection with an explicit focus on TinyML suitability. The dataset includes images of healthy and diseased cocoa pods, leaves, and stems, captured predominantly under uncontrolled field conditions. Images were curated and resized to a low-resolution format to support deployment on resource-constrained edge devices. To demonstrate the practical utility of the dataset, baseline classification experiments were conducted using two lightweight models. The primary contribution of this work is the dataset itself; the models are employed solely to validate its suitability for machine learning and TinyML applications.

The remainder of this paper is structured as follows: Section 2 discusses the TinyML paradigm; Section 3 reviews related datasets; Section 4 describes the methodology used in dataset construction; Section 5 presents baseline experimental results and discussion; and Section 6 concludes the paper with recommendations and future research directions.

The TinyML Paradigm

Tiny Machine Learning (TinyML) refers to the development and deployment of machine learning models on ultra-low-power, resource-constrained edge devices, typically operating within milliwatt-level power budgets (ARM, 2021). Since gaining prominence around 2019, TinyML has attracted significant research interest due to its potential to enable real-time, on-device inference without reliance on cloud connectivity (Ray, 2022).

Article DOI: 10.52589/BJCNIT-NP2MMBZN
DOI URL: https://doi.org/10.52589/BJCNIT-NP2MMBZN

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 1, 2026 (pp. 137-146)
www.abjournals.org

Unlike conventional machine learning workflows that depend heavily on cloud-based computation, TinyML emphasizes efficiency in memory usage, computational complexity, and energy consumption. This paradigm enables reduced latency, improved data privacy, minimal connectivity dependency, and enhanced robustness in remote or bandwidth-limited environments (Banbury et al., 2021). However, these benefits often come at the cost of reduced model capacity and, in some cases, lower predictive accuracy.

In the context of this work, the TinyML paradigm influenced dataset design choices, including image resolution, dataset size, and feature focus. By constraining images to a resolution of 112 × 112 pixels and emphasizing discriminative visual features, CocoaDetectDB is well-suited for training and evaluating models intended for deployment on resource-limited edge devices.

RELATED WORK

Datasets used in computer vision research can broadly be categorized as general-purpose or domain-specific. General-purpose datasets, such as ImageNet (Deng et al., 2009), MS COCO (Lin et al., 2014), and CIFAR-10/100 (Krizhevsky & Hinton, 2009), contain large numbers of images across diverse categories and have played a central role in advancing computer vision research. While these datasets are valuable for benchmarking and pretraining, they are not tailored to agricultural disease detection tasks.

Domain-specific datasets address this limitation by focusing on particular application areas. In agriculture, several datasets have been introduced, including Flowers (M-E Nilsback & Zisserman, 2006; Maria-Elena Nilsback & Zisserman, 2008), VegFru (Hou, Feng, & Wang, 2017), CropDeep (Zheng *et al.*, 2019b), PlantVillage (Hughes & Salathé, 2015), PlantDoc (Singh et al., 2020), Leafsnap (Kumar et al., 2012), and DiaMOS Plant (Fenu & Malloci, 2021). While these datasets have advanced agricultural vision research, they primarily focus on crops other than cocoa and are not explicitly designed for TinyML deployment. A summary of some datasets used in the agricultural domain is shown in Table 1.

To the best of our knowledge, CocoaDetectDB represents the first publicly available cocoa plant disease dataset curated with explicit consideration for TinyML constraints, thereby addressing a critical gap in agricultural computer vision research.

Table 1: Summary of datasets used in the agricultural domain

Dataset	Number of Classes	Total Images
Flowers 102	102	1020
VegFru	70	160,731
CropDeep	31	31,147
PlantVillage	38	54,309
PlantDoc	27	2,598
Leafsnap	185	30,866
DiaMOS plant	4	3,505

Article DOI: 10.52589/BJCNIT-NP2MMBZN
DOI URL: https://doi.org/10.52589/BJCNIT-NP2MMBZN

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 1, 2026 (pp. 137-146)
www.abjournals.org

METHODOLOGY

Data Collection and Preparation

CocoaDetectDB was constructed using a combination of field-collected images and openly accessible public data, following established practices in agricultural dataset development (Li, Zhang, & Wang, 2021). Field data were collected at a cocoa plantation in Akam/Ofutop, Ikom Local Government Area of Cross River State, Nigeria. Images were captured under uncontrolled environmental conditions using consumer-grade devices, including a Samsung Galaxy A12 smartphone, an Infinix S4 smartphone, and a Zinox Edge notepad computer. The specifications of these devices are shown in Table 2.

Images of cocoa pods, leaves, and stems were collected in both healthy and diseased states. Disease identification and annotation were performed by a crop scientist during data collection, ensuring accurate labelling of CBD, CSSVD, and FPR. To augment the dataset, additional images were sourced from publicly available repositories, including Kaggle (Kaggle, 2020) and PlantVillage (PlantVillage, 2023), and selected for research and educational use. CSSVD images were notably scarce due to the disease's geographic restriction to West Africa. Some of the collected data are shown in Figure 1.

Table 2: Specifications of mobile devices used for image capture

Device Name	Manufacturer	Operating System	RAM	Camera Type
Galaxy A12	Samsung	Android 11	4 GB	48 MP
Zinox Edge	Zinox	Windows 10	2 GB	5 MP
Infinix S4	Infinix	Android 9.0 (Pie)	3 GB	13 MP

Data Cleaning and Labelling

A systematic data cleaning process was applied to all collected images. Blurred, poorly focused, or visually uninformative images were discarded. Where necessary, images were cropped to isolate regions of interest containing clear disease symptoms or healthy features. All retained images were resized to a uniform resolution of 112 × 112 pixels with three colour channels (RGB).

The dataset was organized into four mutually exclusive classes:

1. Healthy_pod
2. CBD (Cocoa Black Pod Disease)
3. CSSVD (Cocoa Swollen Shoot Virus Disease)
4. FPR (Frosty Pod Rot)

The entire process of data collection, cleaning, and expert labelling spanned approximately 28 days.

Article DOI: 10.52589/BJCNIT-NP2MMBZN
DOI URL: https://doi.org/10.52589/BJCNIT-NP2MMBZN

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 1, 2026 (pp. 137-146)
www.abjournals.org

Dataset Validation Protocol

To validate the suitability of CocoaDetectDB for machine learning and TinyML applications, baseline classification experiments were conducted. The objective of these experiments was not to propose novel model architectures, but rather to demonstrate that the dataset supports effective automated disease classification under both lightweight and resource-constrained settings.

Two models were employed: MobileNetV2 and a lightweight custom convolutional neural network converted to TensorFlow Lite format. The custom model was quantized to support efficient inference on low-resource devices. Detailed architectural and optimization aspects of the custom model are beyond the scope of this paper and will be presented in a separate study.

All experiments were conducted using Google Colaboratory. Images were resized to 112 × 112 pixels and organized into class-specific directories. The dataset was split into training, validation, and test sets using an 80%:10%:10% ratio. Both models were trained using the same hyperparameters: a batch size of 16, an input depth of 3, and training for 20 epochs.

Figure 1: (a) Samples of Self-Collected Data (b) Data Samples from PlantVillage

(c) Data Samples from Kaggle

(a)

(b)

Article DOI: 10.52589/BJCNIT-NP2MMBZN
DOI URL: https://doi.org/10.52589/BJCNIT-NP2MMBZN



(c)

RESULTS AND DISCUSSION

Dataset Composition

CocoaDetectDB comprises a total of 1,424 images distributed across four classes, as summarized in Table 3. The classes are

- CBD – Cocoa Black Pod Disease (class 0)
- CSSVD – Cocoa Swollen Shoot Virus Disease (class 1)
- FPR – Frosty Pod Rot (class 2)
- Healthy_pod – Healthy Cocoa Pods (class 3)

A portion of the dataset was collected under field conditions in Akam/Ofutop, Ikom Local Government Area of Cross River State, Nigeria. At the same time, the remainder was sourced from publicly available repositories and general web searches. Field-acquired images represent 16% of CBD and 59% of Healthy_pod classes, providing real-world variability in lighting, angle, and background. CSSVD and FPR images were entirely sourced from the internet due to limited geographic distribution.

Table 3: CocoaDetectDB Dataset Distribution

Class	Description	Total Images	Self-Collected	Internet-Sourced	% of Self-Collected
CBD	Cocoa Black Pod Disease	359	58	301	16%
CSSVD	Cocoa Swollen Shoot Virus Disease	224	0	224	0%
FPR	Frosty Pod Rot	234	0	234	0%
Healthy_pod	Healthy Cocoa Pods	607	358	249	59%
Total		1,424	416	1,008	29%

Article DOI: 10.52589/BJCNIT-NP2MMBZN
DOI URL: https://doi.org/10.52589/BJCNIT-NP2MMBZN

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 1, 2026 (pp. 137-146)
www.abjournals.org

The dataset distribution is also visualized in Figure 2, illustrating both the total number of images per class and the proportion of self-collected images.

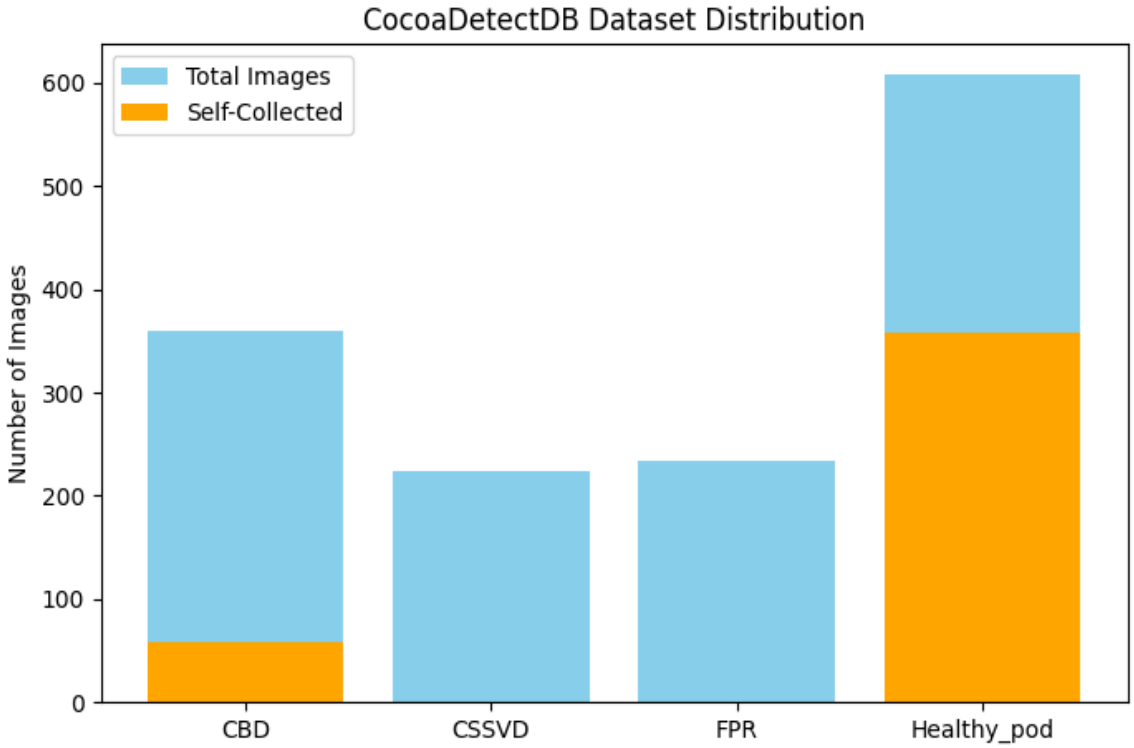
Baseline Validation

To assess the suitability of CocoaDetectDB for machine learning and TinyML applications, baseline classification experiments were conducted. Two models were employed:

1. MobileNetV2 is a widely used lightweight CNN.
2. Custom convolutional neural network, quantized and converted to TensorFlow Lite for edge deployment.

Both models were trained with batch size = 16, image dimensions = 112 × 112 × 3, and 20 epochs, using an 80%:10%:10% split for training, validation, and testing.

Figure 2. Distribution of images in CocoaDetectDB across four classes. Orange bars indicate self-collected images captured under field conditions, while blue bars represent the total number of images per class.



The test accuracies achieved are presented in Table 4.

Table 4: Test Accuracies of Models

Model	Test Accuracy
MobileNetV2	99.13%
Custom TFLite Model	93.75%

Article DOI: 10.52589/BJCNIT-NP2MMBZN
DOI URL: https://doi.org/10.52589/BJCNIT-NP2MMBZN

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 1, 2026 (pp. 137-146)

www.abjournals.org

DISCUSSION

The baseline results demonstrate that CocoaDetectDB provides discriminative visual features sufficient for automated cocoa disease classification, even under resource-constrained TinyML deployment. MobileNetV2 achieved near-perfect accuracy, reflecting both the high-quality labelling and strong separability among the four classes.

The quantized TFLite model achieved a competitive 93.75% accuracy, confirming the dataset's applicability for on-device inference where memory and computational resources are limited. Misclassifications primarily occurred between CBD and FPR, likely due to overlapping visual symptoms on cocoa pods, highlighting realistic challenges in field-acquired datasets.

The combination of field-collected and internet-sourced images ensures that models trained on CocoaDetectDB can generalize across varying environmental conditions, including lighting, background clutter, and image resolution. The relatively high proportion of self-collected images in CBD and Healthy_pod classes further enhance regional relevance for West African cocoa-growing areas.

Overall, these findings validate CocoaDetectDB as a robust, reusable, and publicly accessible dataset, suitable for a wide range of research applications, including:

- TinyML and edge AI deployments
- Transfer learning for disease classification
- Benchmarking lightweight computer vision models
- Practical agricultural monitoring tools for farmers and extension workers

RECOMMENDATIONS

Based on the findings of this study, the following recommendations are proposed:

Research: CocoaDetectDB can serve as a benchmark dataset for TinyML, model compression, and edge AI research in agricultural disease detection.

Practical Deployment: The dataset supports the development of low-cost, on-device cocoa disease monitoring tools for farmers and extension workers.

Policy and Capacity Building: Public investment in region-specific agricultural datasets can significantly enhance AI-driven agricultural innovation, particularly in developing regions.

CONCLUSION

This paper presented CocoaDetectDB, a novel, publicly available image dataset for cocoa plant disease detection designed with TinyML deployment constraints in mind. The dataset captures real-world variability through field-acquired imagery and expert annotation, addressing a critical gap in cocoa-focused agricultural datasets. Baseline validation experiments using MobileNetV2 and a quantized TensorFlow Lite model demonstrate that the dataset supports

Article DOI: 10.52589/BJCNIT-NP2MMBZN
DOI URL: https://doi.org/10.52589/BJCNIT-NP2MMBZN

accurate disease classification under both lightweight and resource-constrained settings. CocoaDetectDB is expected to facilitate future research in cocoa disease detection, edge AI, and precision agriculture, while ongoing work will explore optimized TinyML models in a separate study.

RESOURCES

MobileNetV2 implementation:

https://colab.research.google.com/drive/1bAc4Bz9yF7z5Hd9RawTZgLip7gb7rBDf?usp=sharing

Custom Tensorflow Lite implementation:
https://colab.research.google.com/drive/1aQd8x0ZxlykKNqENCMn0BPYeDTgvYXq5?usp=sharing

CocoaDetectDB Dataset https://drive.google.com/drive/folders/1-Dro6Wq20y6V-0yMutkgnCXMnNAAMW2Z?usp=sharing

REFERENCES

ARM. (2021). Tinyml Brings Ai to Smallest Arm Devices (Vol. 2024).

Banbury, C., Reddi, V. J., Torelli, P., Holleman, J., Jeffries, N., Kiraly, C., Montino, P., Kanter, D., Ahmed, S., & Pau, D. (2021). Mlperf Tiny Benchmark. *arXiv preprint arXiv:2106.07597*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *Imagenet: A Large-Scale Hierarchical Image Database.* Paper presented at the 2009 IEEE conference on computer vision and pattern recognition, Ieee 248-255.

Fenu, G., & Malloci, F. M. (2021). Diamos Plant: A Dataset for Diagnosis and Monitoring Plant Disease. *Agronomy, 11*(11), 2107.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification.* Paper presented at the Proceedings of the IEEE international conference on computer vision 1026-1034.

Hou, S., Feng, Y., & Wang, Z. (2017). *Vegfru: A Domain-Specific Dataset for Fine-Grained Visual Categorization.* Paper presented at the Proceedings of the IEEE international conference on computer vision 541-549.

Hughes, D., & Salathé, M. (2015). An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics. *arXiv preprint arXiv:1511.08060*.

Kaggle. (2020). *Cocoa Diseases*. Retrieved from: https://www.kaggle.com/datasets/serranosebas/enfermedades-cacao-yolov4?resource=download on 15/07/2023

Krizhevsky, A., & Hinton, G. (2009). Learning Multiple Layers of Features from Tiny Images.

Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., & Soares, J. V. (2012). *Leafsnap: A Computer Vision System for Automatic Plant Species Identification.* Paper presented at the European conference on computer vision, Springer 502-516.

Article DOI: 10.52589/BJCNIT-NP2MMBZN
DOI URL: https://doi.org/10.52589/BJCNIT-NP2MMBZN

Li, L., Zhang, S., & Wang, B. (2021). Plant Disease Detection and Classification by Deep Learning—a Review. *IEEE access, 9*, 56683-56698.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). *Microsoft Coco: Common Objects in Context.* Paper presented at the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer 740-755.

Liu, J., & Wang, X. (2021). Plant Diseases and Pests Detection Based on Deep Learning: A Review. *Plant Methods, 17*, 1-18.

Nilsback, M.-E., & Zisserman, A. (2006). *A Visual Vocabulary for Flower Classification.* Paper presented at the 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), IEEE 1447-1454.

Nilsback, M.-E., & Zisserman, A. (2008). *Automated Flower Classification over a Large Number of Classes.* Paper presented at the 2008 Sixth Indian conference on computer vision, graphics & image processing, IEEE 722-729.

PlantVillage. (2023). Cocoa. Retrieved from https://plantvillage.psu.edu/topics/cocoa-cacao/infos on 07/07/2023

Ray, P. P. (2022). A Review on Tinyml: State-of-the-Art and Prospects. *Journal of King Saud University-Computer and Information Sciences, 34*(4), 1595-1623.

Ren, C., Kim, D.-K., & Jeong, D. (2020). A Survey of Deep Learning in Agriculture: Techniques and Their Applications. *Journal of Information Processing Systems, 16*(5), 1015-1033.

Rodriguez, C., Alfaro, O., Paredes, P., Esenarro, D., & Hilario, F. (2021). Machine Learning Techniques in the Detection of Cocoa (Theobroma Cacao L.) Diseases. *Annals of the Romanian Society for Cell Biology, 25*(3), 7732-7741.

Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., & Batra, N. (2020). Plantdoc: A Dataset for Visual Plant Disease Detection *Proceedings of the 7th Acm Ikdd Cods and 25th Comad* (pp. 249-253).

Syamsuri, B., & Kusuma, G. P. (2019). Plant Disease Classification Using Lite Pretrained Deep Convolutional Neural Network on Android Mobile Device. *Int. J. Innov. Technol. Explor. Eng, 9*(2), 2796-2804.

Zheng, Y.-Y., Kong, J.-L., Jin, X.-B., Wang, X.-Y., Su, T.-L., & Zuo, M. (2019). Cropdeep: The Crop Vision Dataset for Deep-Learning-Based Classification and Detection in Precision Agriculture. *Sensors, 19*(5), 1058.

Article DOI: 10.52589/BJCNIT-NP2MMBZN
DOI URL: https://doi.org/10.52589/BJCNIT-NP2MMBZN