IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131006

Crop Disease Classification for Edge Devices: A Quantized MobileNetV2 Approach

Dr. Shilpa Sarvaiya[1], Pranav Dhole[2], Ishika Nandwanshi[3]

Head, Department of Computer Application, Vidyabharti Mahavidyalaya, Amravati, India[1]

Student, Department of Computer Application, Vidyabharti Mahavidyalaya, Amravati, India[2,3]

Abstract: Agriculture plays a critical role in global food security, yet crop diseases continue to cause significant economic losses worldwide. Traditional deep learning models, while achieving high accuracy in disease detection, face substantial limitations when deployed for real-time, in-field applications due to their computational complexity and large memory requirements. This research addresses these challenges by developing a lightweight, quantized model specifically designed for edge device deployment. We propose applying model compression techniques through quantization on MobileNetV2, a lightweight neural network architecture, to create an efficient model suitable for resource-constrained environments. Our methodology involves comprehensive comparison of Post-Training Quantization (PTQ) and Dynamic Range Quantization (DRQ) techniques applied to rice leaf disease classification. The results demonstrate a significant reduction in model size from approximately 9 MB to 2.5 MB while maintaining acceptable accuracy levels. The DRQ model achieved 92.23% accuracy with an F1-score of 0.9212, compared to the original model's 94% accuracy, representing a minimal 1.77% accuracy trade-off for a 72% size reduction. These findings highlight the practical viability of quantized models for automated disease detection systems in precision agriculture, enabling real-time deployment on smartphones and embedded devices for farmers in remote locations.

Keywords: Crop disease classification, edge computing, model quantization, MobileNetV2, precision agriculture, deep learning compression.

I. INTRODUCTION

Agriculture remains the backbone of global food security, supporting billions of people worldwide and contributing significantly to economic development in many countries. However, plant diseases pose a persistent threat to crop productivity, causing annual losses estimated at 20-40% of global crop production, translating to billions of dollars in economic impact [1]. Traditional disease identification methods rely heavily on expert knowledge and visual inspection, which are time-consuming, subjective, and often unavailable in remote agricultural areas.

The advent of deep learning has revolutionized plant disease detection, with Convolutional Neural Networks (CNNs) achieving remarkable accuracy in identifying various crop diseases from digital images [2]. However, these high-performing models typically require substantial computational resources, large memory footprints, and significant power consumption, making them impractical for deployment on edge devices such as smartphones, tablets, or embedded systems commonly used in agricultural settings.

The challenge of deploying sophisticated AI models in resource-constrained environments has led to increased inter- est in model compression techniques. Edge computing in agriculture offers numerous advantages, including real-time processing, reduced dependency on internet connectivity, lower latency, and enhanced privacy for farmers' data. However, the computational limitations of edge devices necessitate the development of lightweight models that can maintain acceptable accuracy while operating within strict resource constraints.

This research addresses the critical need for efficient crop disease classification models suitable for edge deployment. Our primary objective is to develop a highly efficient and accurate model for rice leaf disease classification by leveraging quantization techniques to significantly reduce model size and improve inference performance on edge devices. Specifically, we aim to: (1) implement and compare different quantization approaches on a lightweight architecture, (2) analyze the trade-offs between model size, accuracy, and inference speed, and (3) demonstrate the practical viability of quantized models for real-world agricultural applications.

The remainder of this paper is organized as follows: Section II presents a comprehensive literature review of existing approaches to plant disease detection and model compression techniques. Section III details our methodology, including

DOI: 10.17148/IJIREEICE.2025.131006

dataset description, model architecture, and quantization procedures. Section IV presents experimental results, followed by analysis and discussion in Section V. Finally, Section VI concludes the paper and outlines future research directions.

II. LITERATURE REVIEW

2.1 Traditional Methods for Plant Disease Detection

Early approaches to automated plant disease detection relied primarily on traditional computer vision techniques and handcrafted feature extraction methods. These systems typically employed color-based segmentation, texture analysis, and morphological operations to identify diseased regions in plant images [3]. While these methods provided interpretable results and required minimal computational resources, they struggled with complex backgrounds, varying lighting con- ditions, and the subtle visual differences between disease symptoms.

Feature extraction techniques such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Scale-Invariant Feature Transform (SIFT) were commonly employed in conjunction with traditional machine learning classifiers like Support Vector Machines (SVM) and Random Forest [4]. However, these approaches required extensive domain expertise for feature engineering and often failed to generalize across different crops, diseases, or environmental conditions.

2.2 Deep Learning for Plant Disease Classification

The introduction of deep learning, particularly CNNs, marked a significant advancement in plant disease detection accuracy. Pioneering work by Mohanty et al. [2] demonstrated that deep CNNs could achieve over 99% accuracy on the PlantVillage dataset, significantly outperforming traditional methods. Subsequent research has explored various CNN architectures, including ResNet, DenseNet, and Inception networks, consistently achieving high classification accuracy across multiple crops and diseases [5].

Transfer learning has emerged as a particularly effective approach, allowing researchers to leverage pre-trained models from large-scale datasets like ImageNet and fine-tune them for specific plant disease classification tasks [6]. This approach has proven especially valuable when dealing with limited agricultural datasets, enabling high accuracy with reduced training time and computational requirements.

Despite their success, these high-capacity models typically range from 25 MB to over 500 MB in size and require substantial computational resources for inference, making them unsuitable for deployment on resource-constrained edge devices commonly used in agricultural applications.

2.3 Lightweight Architectures for Mobile and Edge Computing

Recognizing the need for efficient models, researchers have developed lightweight CNN architectures specifically designed for mobile and edge applications. MobileNet [7] introduced depthwise separable convolutions, which factorize standard convolutions into depthwise and pointwise operations, significantly reducing computational complexity while maintaining reasonable accuracy.

MobileNetV2 [8] further improved upon the original design by incorporating inverted residuals and linear bottlenecks, achieving better accuracy-efficiency trade-offs. EfficientNet [9] proposed compound scaling of network depth, width, and resolution, demonstrating that careful scaling can achieve superior performance with fewer parameters.

These architectures have shown promise for agricultural applications, with several studies demonstrating their effectiveness for crop disease detection while maintaining relatively small model sizes [10]. However, even these lightweight models often exceed the memory and computational constraints of low-end edge devices.

2.4 Model Compression and Quantization Techniques

Model compression encompasses various techniques aimed at reducing the size and computational requirements of neural networks while preserving their performance. Quantization, one of the most effective compression methods, reduces the precision of model weights and activations from 32-bit floating-point to lower-precision representations, typically 8-bit integers [11].

Post-Training Quantization (PTQ) applies quantization to a pre-trained model without requiring retraining, making it a convenient and fast compression method [12]. However, PTQ may result in significant accuracy degradation, particularly for models with complex architectures or when using aggressive quantization schemes.

Dynamic Range Quantization (DRQ) represents a more sophisticated approach that dynamically determines the optimal quantization parameters during the conversion process, often achieving better accuracy preservation compared to static quantization methods [13].

Quantization-Aware Training (QAT) incorporates quantization effects during the training process, allowing the model to adapt to the reduced precision and typically achieving the best accuracy-size trade-offs [11]. However, QAT requires access to the training dataset and significantly more computational resources.

Several studies have applied quantization techniques to agricultural applications, demonstrating promising results for

DOI: 10.17148/IJIREEICE.2025.131006

crop disease detection on mobile devices [14]. However, comprehensive comparisons of different quantization approaches specifically for agricultural edge deployment remain limited.

III. METHODOLOGY

3.1 Dataset

This research utilizes the "Rice Leafs Disease Dataset" available on Kaggle, which contains high-quality images of rice leaves affected by various diseases. The dataset comprises 5,932 images distributed across six classes: bacterial leaf blight, brown spot, healthy leaves, leaf blast, leaf scald, and narrow brown spot. Each class contains between 900-1,000 images, providing a relatively balanced dataset for training and evaluation.

The images were captured under controlled conditions with consistent lighting and backgrounds, featuring close-up views of rice leaves with clear disease symptoms. Image resolution varies between 256x256 and 512x512 pixels, with all images converted to 224x224 pixels to match the input requirements of MobileNetV2. The dataset was randomly split into training (70%), validation (15%), and testing (15%) sets, ensuring representative distribution across all disease classes.

3.2 Base Model Architecture

We selected MobileNetV2 as the base architecture due to its proven efficiency and suitability for mobile deployment. MobileNetV2 employs several key innovations that make it particularly suitable for edge applications:

- Depthwise separable convolutions that reduce computational complexity
- Inverted residual blocks with linear bottlenecks
- Efficient use of ReLU6 activation functions
- Overall parameter count of approximately 3.4 million

The model implementation utilized transfer learning, initializing with weights pre-trained on ImageNet. The final classification layer was replaced with a dense layer containing six neurons (corresponding to the six disease classes) with softmax activation. The architecture consists of the MobileNetV2 base model with frozen weights, followed by global average pooling, dropout layer (0.2), and a final dense classification layer.

3.3 Model Training

The baseline full-precision model was trained using the following configuration:

- Optimizer: Adam with learning rate 0.001
- Loss function: Categorical crossentropy
- Batch size: 32
- Training epochs: 50 with early stopping
- Hardware: NVIDIA Tesla V100 GPU

Data augmentation techniques were applied during training to improve generalization, including random rotation (±15 degrees), width and height shifts (±0.1), horizontal flipping, and zoom range (0.1). The model achieved convergence after 35 epochs with a final validation accuracy of 94.2%.

3.4 Quantization Implementation

Two quantization techniques were implemented and compared:

3.1.1Post-Training Quantization (PTQ):

This approach converts the trained model to TensorFlow Lite format with 8-bit integer quantization applied to both weights and activations. The conversion process applies uniform quantization parameters across the entire model, optimizing for storage efficiency while maintaining compatibility with edge hardware accelerators.

3.1.2 Dynamic Range Quantization (DRQ):

This method applies quantization dynamically, determining optimal quantization parameters during conversion through analysis of activation distributions. A representative dataset is used during conversion to calibrate the quantization parameters for each layer, enabling more precise quantization that better preserves model accuracy.

Both quantized models were evaluated on the test dataset using TensorFlow Lite interpreter to ensure accurate performance measurement under deployment conditions.

IV. RESULTS

4.1 Model Performance Comparison

Table I presents a comprehensive comparison of the three models across key performance metrics. The results demonstrate the effectiveness of quantization in achieving significant model size reduction while maintaining acceptable accuracy levels.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131006

Table 1: Comparison of Model Performance Metrics

Model	Accuracy (%)	Size (MB)	F1-Score (Macro)
Full-Precision	94.00	9.0	0.9374
PTQ Quantized	83.00	2.5	0.8245
DRQ Quantized	92.23	2.59	0.9212

The full-precision baseline model achieved 94% accuracy with a model size of approximately 9 MB. Both quantization techniques successfully reduced the model size to approximately 2.5 MB, representing a 72% reduction in storage requirements. However, the quantization approaches showed significantly different accuracy preservation characteristics.

4.2 Detailed Performance Analysis

4.1.1 Post-Training Quantization Results:

The PTQ model showed substantial accuracy degradation, achieving 83% accuracy compared to the baseline's 94%. Table II presents the detailed per-class performance metrics, revealing significant variations in quantization impact across different disease types.

Table 2: Post-Training Quantization (PTQ) - Per-Class Performance Metrics

Disease Class	Precision	Recall	F1-Score
Bacterial Leaf Blight	0.75	1.00	0.85
Brown Spot	0.89	0.67	0.77
Healthy	0.93	0.91	0.92
Leaf Blast	0.64	0.89	0.74
Leaf Scald	0.95	0.94	0.95
Narrow Brown Spot	1.00	0.56	0.72
Macro Average	0.86	0.83	0.8245

The confusion matrix analysis revealed that PTQ particularly struggled with distinguishing between leaf blast and brown spot, with frequent misclassifications between these visually similar conditions. Notably, the narrow brown spot class showed perfect precision but low recall, indicating conservative classification behavior.

4.1.2 Dynamic Range Quantization Results:

The DRQ model demonstrated superior performance, maintaining 92.23% accuracy with only a 1.77% degradation from the baseline. Table III shows the per-class metrics, which consistently demonstrate higher performance across all disease categories compared to PTQ.

Table 3: Dynamic Range Quantization (DRQ) - Per-Class Performance Metrics

Disease Class	Precision	Recall	F1-Score
Bacterial Leaf Blight	0.97	1.00	0.98
Brown Spot	0.95	0.70	0.81
Healthy	0.92	0.95	0.94
Leaf Blast	0.76	0.91	0.83
Leaf Scald	0.98	1.00	0.99
Narrow Brown Spot	0.99	0.97	0.98
Macro Average	0.93	0.92	0.9212

The DRQ model maintained high precision and recall across most classes, with particularly strong performance in detecting leaf scald and narrow brown spot. The macro-averaged F1-score of 0.9212 indicates robust performance across all disease categories. Compared to PTQ, DRQ shows significant improvements in recall for narrow brown spot (0.97 vs 0.56) and precision for leaf blast (0.76 vs 0.64).

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131006

4.3 Inference Performance

Inference time measurements were conducted on a Raspberry Pi 4 Model B to evaluate real-world edge deployment performance. Table IV summarizes the inference speed comparison across all three models.

Table 4: Inference Performance Comparison on Raspberry Pi 4 Model B

Model	Inference Time (seconds)	Speedup Factor
Full-Precision	2.3	1.0x (baseline)
PTQ Quantized	0.8	2.9x
DRQ Quantized	0.9	2.6x

The quantized models demonstrated significant speedup compared to the full-precision model, with PTQ achieving the fastest inference time of 0.8 seconds per image and DRQ following closely at 0.9 seconds per image. These results confirm the practical benefits of quantization for edge deployment, enabling near real-time inference on resource-constrained devices.

V. ANALYSIS AND DISCUSSION

5.1 Impact of Quantization on Model Efficiency

The experimental results clearly demonstrate the effectiveness of quantization in creating deployable models for edge devices. The reduction from 9 MB to approximately 2.5 MB represents a critical improvement for applications targeting smartphones, tablets, and embedded systems with limited storage capacity. This size reduction enables deployment scenarios previously impractical with full-precision models, including offline operation in areas with limited internet connectivity.

The storage efficiency gains translate directly to practical benefits for agricultural applications. A 2.5 MB model can be easily embedded within mobile applications, updated over cellular networks with minimal data costs, and stored on devices with limited memory. This accessibility is particularly important for smallholder farmers in developing countries who may rely on low-cost smartphones for agricultural guidance.

5.2 Comparative Analysis of Quantization Techniques

The performance comparison between PTQ and DRQ reveals significant differences in their effectiveness for agricultural image classification. The DRQ model's superior performance (92.23% vs. 83% accuracy) can be attributed to its more sophisticated approach to quantization parameter selection.

PTQ applies uniform quantization parameters across the entire model, which may not optimally represent the diverse feature distributions learned by different layers. This limitation becomes particularly problematic for agricultural images, where subtle visual differences between disease symptoms require precise feature representation. The 11-point accuracy gap between PTQ and DRQ methods highlights this limitation.

DRQ's dynamic parameter selection allows for layer-specific optimization, better preserving the discriminative features crucial for disease classification. The per-class performance analysis reveals that DRQ consistently outperforms PTQ across most disease categories, with particularly notable improvements in challenging cases like leaf blast detection (F1-score: 0.83 vs. 0.74).

5.3 Practical Implications for Agricultural Deployment

The DRQ model's combination of 92.23% accuracy, 2.59 MB size, and 2.6x inference speedup makes it highly suitable for practical agricultural applications. The minimal accuracy loss (1.77%) represents an acceptable trade-off considering the substantial efficiency gains. This performance level exceeds many human experts' consistency in disease identification, particularly for early-stage symptoms or visually similar conditions.

The real-world deployment scenario analysis suggests that the DRQ model can enable several practical applications:

- Mobile applications for farmers providing instant disease diagnosis
- Integration with drone-based crop monitoring systems
- Embedded systems for continuous field monitoring
- Educational tools for agricultural extension services

The inference speed of 0.9 seconds per image on Raspberry Pi 4 enables near real-time processing for individual image analysis while remaining suitable for batch processing of multiple images captured during field surveys.

5.4 Limitations and Challenges

Despite the promising results, several limitations must be acknowledged. The dataset's controlled conditions may not

DOI: 10.17148/IJIREEICE.2025.131006

fully represent the variability encountered in real-world agricultural settings, including diverse lighting conditions, camera angles, and image quality variations. The performance degradation observed with aggressive quantization (PTQ) indicates that further compression may require more sophisticated approaches such as Quantization-Aware Training.

The current evaluation focuses on a single crop (rice) and may not generalize to other crops with different visual characteristics or disease symptoms. Additionally, the model's performance on images captured with varying camera sensors, resolutions, and preprocessing pipelines requires further investigation.

5.5 Future Research Directions

Several avenues for future research emerge from this work. Quantization-Aware Training represents a promising approach for further improving the accuracy-efficiency trade-off by incorporating quantization effects during the training process. Hybrid quantization schemes that apply different precision levels to different model layers could potentially achieve better performance than uniform quantization approaches.

The integration of knowledge distillation with quantization could enable the development of even smaller models while maintaining high accuracy. Additionally, investigating neural architecture search (NAS) techniques specifically optimized for quantized models could lead to architectures better suited for compression.

From an application perspective, expanding the evaluation to include multiple crops, diverse environmental conditions, and different camera sensors would provide valuable insights into the model's generalization capabilities. The development of continual learning approaches that allow models to adapt to new diseases or environmental conditions while maintaining deployment efficiency represents another important research direction.

VI. CONCLUSION

This research demonstrates the practical viability of quantized deep learning models for crop disease classification on edge devices. Through comprehensive evaluation of MobileNetV2 with different quantization approaches, we have shown that model compression can achieve substantial efficiency improvements while maintaining acceptable accuracy for agricultural applications.

The key findings of this study include: (1) quantization successfully reduced model size by 72% (from 9 MB to 2.5 MB) while achieving significant inference speedup on edge devices, (2) Dynamic Range Quantization outperformed Post-Training Quantization, maintaining 92.23% accuracy compared to the original model's 94%, and (3) the quantized models demonstrate practical deployment viability with inference times suitable for real-world agricultural applications.

The DRQ approach emerges as the recommended quantization technique for agricultural edge applications, offering the optimal balance between model size, accuracy, and inference speed. The minimal accuracy degradation (1.77%) represents an acceptable trade-off for the substantial efficiency gains, making automated rice disease detection feasible on resource-constrained devices commonly available to farmers.

These findings have significant implications for precision agriculture, particularly in developing countries where access to agricultural expertise may be limited. The deployment of efficient disease detection models on mobile devices can democratize access to advanced agricultural diagnostics, potentially reducing crop losses and improving food security.

Future work should focus on expanding the evaluation to multiple crops, implementing Quantization-Aware Training approaches, and developing more sophisticated compression techniques that can achieve even better accuracy-efficiency trade-offs. The integration of these models with complete agricultural decision support systems represents an important step toward realizing the full potential of AI in sustainable agriculture.

REFERENCES

[1]. S. Savary et al., "The global burden of pathogens and pests on major food crops," *Nature Ecology & Evolution*, vol. 3, no. 3, pp. 430-439, 2019.

[2]. S. P. Mohanty, D. P. Hughes, and M. Salathe´, "Using deep learning for image-based plant disease detection," *Fron- tiers in Plant Science*, vol. 7, p. 1419, 2016.

[3]. C. H. Bock et al., "Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging," *Critical Reviews in Plant Sciences*, vol. 29, no. 2, pp. 59-107, 2010.

[4]. R. Pydipati, T. F. Burks, and W. S. Lee, "Identification of citrus disease using color texture features and discriminant analysis," *Computers and Electronics in Agriculture*, vol. 52, no. 1-2, pp. 49-59, 2006.

[5]. E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture*, vol. 161, pp. 272-279, 2019.

[6]. M. Brahimi, K. Boukhalfa, and A. Moussaoui, "Deep learning for tomato diseases: classification and symptoms visualization," *Applied Artificial Intelligence*, vol. 31, no. 4, pp. 299-315, 2017.

[7]. A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications,"

DOI: 10.17148/IJIREEICE.2025.131006

arXiv preprint arXiv:1704.04861, 2017.

[8]. M. Sandler et al., "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510-4520.

[9]. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 6105-6114.

[10]. Q. H. Cap et al., "LeafGAN: An effective data augmentation method for practical plant disease diagnosis," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 2, pp. 1258-1267, 2020.

[11]. B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2704-2713.

[12]. R. Banner, Y. Nahshan, and D. Soudry, "Post training 4-bit quantization of convolutional networks for rapid- deployment," in *Advances in Neural Information Processing Systems*, 2019, pp. 7950-7958.

[13]. R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," arXiv preprint arXiv:1806.08342, 2018.

[14]. J. Liu and X. Wang, "Plant diseases recognition based on image processing technology," *Journal of Electrical and Computer Engineering*, vol. 2020, Article ID 6070284, 2020.