

How Do Secondary Plant Health Symptoms Degrade Deep Learning Performance in Black Pepper Leaf Analysis?

¹ Nung Kion Lee
Faculty of Computer Science
& Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Sarawak
nklee@unimas.my

² Olivia Ching Hui Chen
Faculty of Cognitive Sciences
& Human Development
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
olivia.ch.chen@gmail.com

^{3,*} Chih How Bong
Faculty of Computer Science
& Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
chbong@unimas.my

Abstract—Black pepper is an important agricultural commodity in Sarawak and contributes to the livelihoods of local farming communities and the national economy. However, black pepper plants are susceptible to diseases and nutrient deficiencies that can substantially reduce plant growth and crop yield. Timely intervention is constrained by limited personnel and the logistical challenges associated with assessing remote farms. Machine-learning-based diagnostic systems offer a feasible means of addressing these constraints because of their effectiveness in identifying crop health conditions. Nevertheless, the presence of secondary disease or nutrient-deficiency symptoms in leaf images may complicate classification. This study established a diagnostic benchmark by evaluating six Convolutional Neural Network (CNN) models on datasets containing single-symptom and hybrid-symptom images. EfficientNet-B0 achieved the best overall classification performance, whereas ResNet-V2-50 was the most resilient to the introduction of hybrid images. The custom CNN also achieved performance comparable to that of the established architectures. Across the evaluated models, the inclusion of secondary symptoms generally reduced classification performance. These findings indicate a need for methods that mitigate the effects of secondary symptoms in leaf images to improve the reliability of machine-learning-based plant health diagnosis.

Index Terms—Convolutional Neural Networks, Agriculture, Plant Diseases, Black Pepper

I. INTRODUCTION

Black pepper (*Piper nigrum* L.) is an important agricultural commodity in Sarawak, Malaysia, on the island of Borneo. However, black pepper plants are vulnerable to diseases and nutrient deficiencies that can substantially reduce plant growth and crop yield. Farmers typically rely on visual inspection to diagnose these conditions. This manual process can be time-consuming and ambiguous, often requiring expert assistance and potentially delaying treatment and causing substantial financial losses.

Computer-assisted identification of plant health conditions from images, particularly through Convolutional Neural Networks (CNNs), provides a means of automating plant diag-

nosis. Images that display symptoms of diseases or nutrient deficiencies can be collected to train computational models [1]. Leaf images are commonly used because leaves often exhibit visible indicators of plant health. CNNs can analyse these images to identify complex visual patterns associated with specific conditions. Although deep learning has demonstrated strong performance in crop diagnosis, existing public datasets and computational tools rarely address black pepper. Therefore, CNN models developed specifically for black pepper could provide farmers with more rapid diagnostic support.

Black pepper plants may exhibit multiple co-occurring health conditions, such as red rust disease and potassium deficiency. However, most computational studies focus on identifying individual diseases or deficiencies and do not consider overlapping symptoms. Introducing a hybrid class may reduce the predictive performance of standard CNN models because symptoms shared across classes can complicate feature learning. Evaluating the effects of such hybrid symptoms is therefore important for the development of diagnostic systems suitable for real-world crop management.

Figure 1 shows a leaf exhibiting symptoms of both sooty mould and red rust.



Fig. 1. A black pepper leaf exhibiting co-occurring symptoms of sooty mould and red rust, illustrating a hybrid-symptom image considered in this study (source: authors).

*Corresponding author

This study investigates how secondary symptoms of disease or nutrient deficiency affect the classification performance of deep neural networks. Six CNN architectures are evaluated using black pepper leaf images collected from farms in Sarawak.

The remainder of this paper is organised as follows. Section II reviews related applications of CNNs to plant disease classification. Section III describes the datasets, preprocessing procedures, neural network architectures, experimental settings, and performance metrics. Section IV presents and discusses the experimental results.

II. RELATED WORK

Convolutional Neural Networks (CNNs) have emerged as effective computational tools for diagnosing plant diseases. Their ability to learn relevant visual features automatically has enabled them to outperform traditional machine-learning approaches that rely on hand-crafted features. Previous studies have demonstrated the effectiveness of CNNs in identifying various plant health conditions. For example, [2] developed MobInc-Net, a lightweight Inception-based network for recognising and detecting rice diseases. When evaluated on a locally collected rice image dataset and the public PlantVillage dataset, MobInc-Net achieved a peak accuracy of 99.21% on the local dataset, demonstrating the potential of lightweight models. To address the limitations of small datasets, [3] used a Super-Resolution Generative Adversarial Network (SRGAN) to augment rice leaf disease images before classification. DenseNet169 and MobileNetV2 each achieved an accuracy of 94.30% on the augmented dataset, compared with 74.86% for VGG16.

Other studies have evaluated pretrained and computationally efficient CNN models on diverse datasets. [4] evaluated several architectures using rice crop images collected from farms in Tamil Nadu, India. ResNet152 achieved the highest accuracy (98.66%), followed by MobileNet (98.59%); the evaluated DenseNet configurations achieved accuracies ranging from 94.37% to 96.55%. Similarly, [5] compared pretrained architectures using a rice disease image dataset obtained from Kaggle. InceptionV3 achieved an accuracy of 93.91%, followed by ResNet-50 at 90.00% and VGG-16 at 86.50%. Collectively, these studies indicate that lightweight, pretrained, and augmentation-based CNN approaches can provide accurate plant health classification. However, they primarily address single-condition images and do not examine how co-occurring symptoms affect model performance.

III. METHOD

To evaluate the effect of secondary symptoms on CNN performance, two datasets were prepared. The SINGLE dataset contained images with one labelled health condition, whereas the HYBRID dataset added a class containing images with both potassium deficiency and red rust. The six CNN models were first evaluated on the SINGLE dataset and then evaluated on the HYBRID dataset under the same experimental conditions.

A. Dataset Preparation

The leaf image dataset was collected from black pepper farms in several divisions of Sarawak, Malaysia: Betong, Bintangor, Bintulu, Serian, Sibul, and Sri Aman. Leaves were sampled randomly from black pepper plants and photographed against a plain background under controlled conditions. The images were captured using digital cameras with a resolution of at least 20 megapixels.

The SINGLE dataset comprised 1,041 images distributed across 11 classes: healthy leaves; three disease classes (red rust, sooty mould, and viral disease); and seven nutrient-related classes (acid soil, calcium deficiency, iron deficiency, magnesium deficiency, manganese deficiency, nitrogen deficiency, and potassium deficiency). This dataset was previously used in [6]. An additional class of 52 images showing both potassium deficiency and red rust was then added to form the 12-class HYBRID dataset. Red rust was selected as the secondary symptom because it commonly co-occurs with other disease or nutrient-deficiency symptoms. Figure 2 shows an example of a hybrid image.

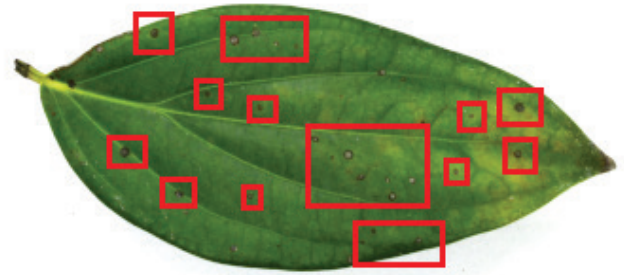


Fig. 2. A leaf exhibiting potassium deficiency with red rust as a secondary symptom.

The SINGLE and HYBRID datasets underwent the same preprocessing procedure (see Fig. 3). First, the leaf regions were segmented from the background using Otsu thresholding [7] and adaptive thresholding implemented in OpenCV. Images in which shadows or damaged leaf areas prevented automatic separation were segmented manually using Adobe Photoshop. The images were then cropped horizontally and vertically to remove excess background and resized to 224×224 pixels. Data augmentation was performed by rotating the images by 45° , 90° , 135° , 180° , 225° , and 270° to improve model robustness and generalisation to unseen data. Finally, the pixel values were normalised to a common range to facilitate convergence during training. After augmentation, the SINGLE dataset contained 4,164 images, and the HYBRID dataset contained 4,372 images. Table I summarises the numbers of original and augmented images.

B. Experimental Setup

The experiments comprised two stages: (1) classification using the 11-class SINGLE dataset and (2) classification using

TABLE I

NUMBERS OF ORIGINAL AND AUGMENTED LEAF IMAGES USED TO TRAIN AND EVALUATE THE CNN MODELS. THE AUGMENTED TOTALS INCLUDE THE ORIGINAL IMAGES.

Category	Class	Total	Total after augmentation
Healthy	Healthy	142	568
Diseases	Red rust	40	496
	Sooty mould	124	160
	Viral disease	274	1096
Nutrient imbalance	Acid soil	24	96
	Calcium deficiency	42	276
	Iron deficiency	45	168
	Magnesium deficiency	48	192
	Manganese deficiency	69	180
	Nitrogen deficiency	108	432
	Potassium deficiency	125	500
Hybrid	Red rust and potassium deficiency	52	208
	Total	1093	4372

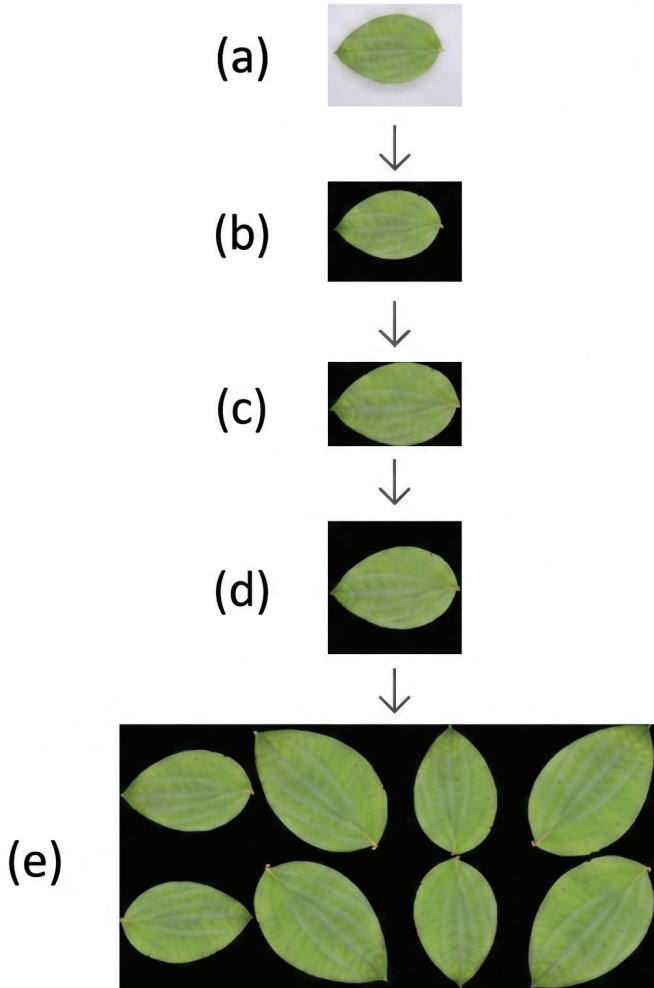


Fig. 3. Image preprocessing pipeline: (a) original image; (b) segmentation of the leaf region from the background; (c) removal of horizontal and vertical regions containing no leaf data; (d) cropping to a square shape and resizing to 224×224 pixels; and (e) data augmentation through image rotation.

the 12-class HYBRID dataset.

Five established CNN architectures were selected: DenseNet121 [8], EfficientNet-B0 [9], MobileNet-V2 [10], ResNet-V2-50 [11], and Inception-V3 [12]. In addition, 28 custom CNN architectures with different layer configurations were evaluated. Based on their performance on the SINGLE dataset, the best custom architecture was selected. This architecture contained five repeated blocks, each consisting of one convolutional layer followed by one pooling layer; it is subsequently referred to as Custom-5-1. Further details of the experimental setup and custom CNN configurations are provided in [6].

C. Performance Metrics

The neural networks were evaluated using five-fold cross-validation. The performance metrics were accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

where TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively. The AUC-ROC summarises classification performance across all possible decision thresholds [13]. It therefore evaluates a model's ability to distinguish among classes without relying on a single threshold or a specific trade-off between sensitivity and specificity.

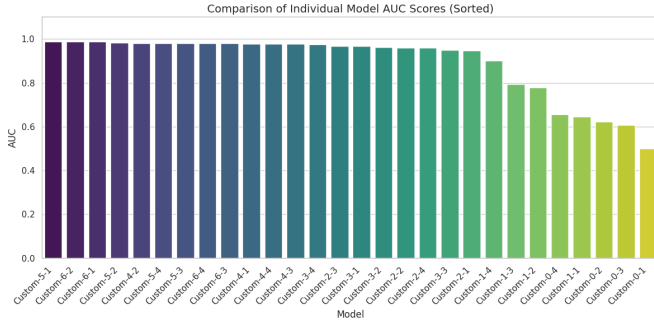


Fig. 4. AUC values of the 28 custom CNN models. The notation $x-y$ denotes x repeated blocks, each containing y convolutional layers followed by one pooling layer.

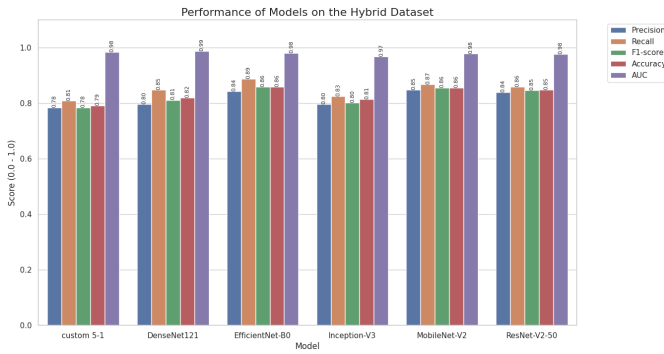


Fig. 5. Performance of the six CNN models on the HYBRID dataset, averaged across the five cross-validation folds.

IV. RESULTS

Detailed evaluation results for the SINGLE dataset are reported in [6]. In summary, EfficientNet-B0 was the strongest baseline classifier for individual black pepper diseases and nutrient deficiencies, achieving the highest accuracy, precision, and recall.

Among the 28 custom CNN architectures, Custom-5-1 achieved the best performance. It comprised five repeated blocks, each with one convolutional layer followed by one pooling layer, and used batch normalisation, ReLU activation, dropout, and two final fully connected layers. Figure 4 presents the AUC values of the 28 custom CNN models.

TABLE II
HIGHEST PERFORMANCE VALUES AMONG THE EVALUATED ARCHITECTURES ON THE HYBRID DATASET.

Metric	Best-performing model(s)	Score
Accuracy	EfficientNet-B0	0.8600
Precision	MobileNet-V2	0.8480
Recall	EfficientNet-B0	0.8880
F1-score	EfficientNet-B0 / MobileNet-V2	0.8560
AUC	DenseNet121	0.9871

A. HYBRID Dataset

Figure 5 compares the performance of Custom-5-1, DenseNet121, EfficientNet-B0, Inception-V3, MobileNet-V2,

and ResNet-V2-50 on the HYBRID dataset. The results reveal the following performance trends.

1) EfficientNet-B0 sensitivity and accuracy:

EfficientNet-B0 achieved the highest recall (0.888) and accuracy (0.860) on the HYBRID dataset. Its compound-scaling approach, which balances network depth, width, and input resolution, may help it capture the heterogeneous features in hybrid images. The high recall indicates that EfficientNet-B0 produced fewer false negatives than the other evaluated models.

2) Discriminative performance measured by AUC:

All six architectures achieved AUC values above 0.968. Custom-5-1 and DenseNet121 achieved AUC values of 0.983 and 0.987, respectively. Thus, despite differences in precision and recall at the selected decision thresholds, all models retained a strong overall ability to distinguish among classes.

3) Performance of the lightweight architecture:

MobileNet-V2 achieved an F1-score of 0.856 and precision of 0.848, exceeding the precision of the larger Inception-V3 architecture (0.796). These results suggest that MobileNet-V2's inverted residual structure and depthwise separable convolutions provide effective feature extraction for this dataset and may offer a favourable balance between computational complexity and predictive performance.

4) Precision of the custom and Inception architectures:

Custom-5-1 and Inception-V3 achieved the lowest precision values, at 0.784 and 0.796, respectively, indicating that they produced more false positives than the other models. For Inception-V3, the contrast between its relatively low precision and high AUC suggests that recalibrating the decision threshold may improve performance on the HYBRID dataset.

5) Stability of ResNet-V2-50:

ResNet-V2-50 produced balanced results across all five metrics, including precision of 0.840, recall of 0.860, and accuracy of 0.848. Its residual mappings may support stable gradient flow and reduce the performance degradation associated with deeper networks.

Table II identifies the best-performing model for each metric.

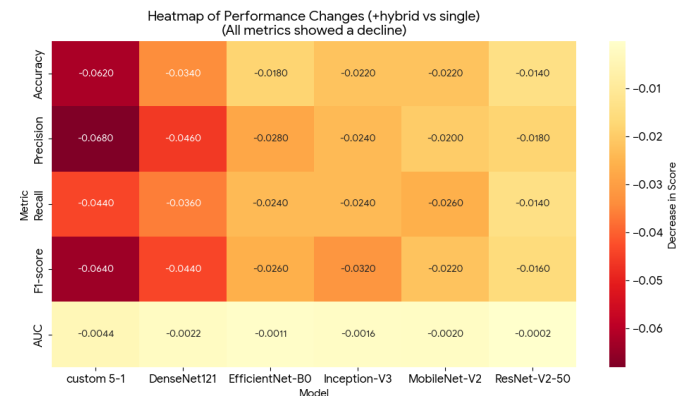


Fig. 6. Performance changes for the six CNN models on the HYBRID dataset relative to the SINGLE dataset. Values indicate percentage changes.

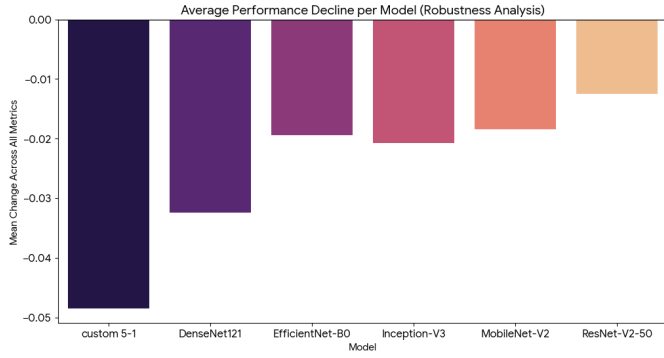


Fig. 7. Average changes in the five performance metrics (accuracy, precision, recall, F1-score, and AUC) for the six CNN models.

Figure 6 compares the performance of the models on the SINGLE and HYBRID datasets. Overall, adding images with secondary symptoms reduced the average performance of every model. The decrease did not exceed 7% for any metric, and the largest reductions generally occurred in precision and F1-score. For Custom-5-1, precision and F1-score decreased by 6.8% and 6.4%, respectively; for DenseNet121, they decreased by 4.6% and 4.4%, respectively. The hybrid images also increased the number of false positives across the models, likely because overlapping symptoms made the classes more difficult to distinguish.

Custom-5-1 exhibited the largest overall reductions across the performance metrics (Fig. 7). Its test accuracy decreased by 6.20 percentage points, from 85.40% to 79.20%, and its precision decreased by 6.80 percentage points. These results suggest that custom architectures may require more careful hyperparameter selection and structural configuration when applied to more complex datasets.

Conversely, ResNet-V2-50 was the most resilient to the introduction of hybrid images, with the smallest decreases in test accuracy (1.40 percentage points) and test precision (1.80 percentage points). Figure 8 compares the confusion matrices produced by ResNet-V2-50 for the SINGLE and HYBRID datasets.

The class-level analysis showed that the AUC for red rust underwent the largest decrease across the models, declining by 1.40 percentage points for Inception-V3, 1.20 percentage points for Custom-5-1, and 1.00 percentage point for DenseNet121. In contrast, adding the hybrid class increased the AUC for some conditions, including sooty mould and magnesium deficiency, in several models.

Figure 9 shows the class-level AUC values obtained on the HYBRID dataset. The inclusion of images with a secondary symptom produced mixed changes in AUC. Of the 66 values compared across the 11 shared classes and six models, 23 were unchanged (34.85%), 14 increased (21.21%), and 29 decreased (43.94%). The decreases ranged from 0.2 to 1.2 percentage points. DenseNet121 produced unchanged AUC values for 6 of the 11 classes, suggesting relative robustness at the class level. Despite the overall reduction in performance, EfficientNet-B0

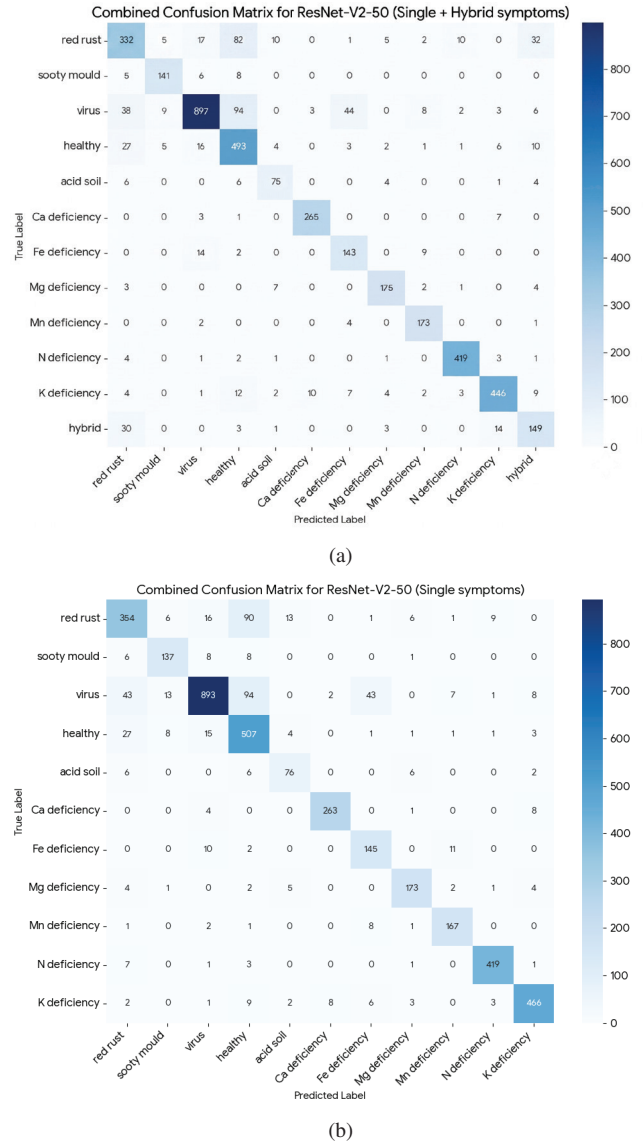


Fig. 8. Combined confusion matrices for ResNet-V2-50 across five test sets: (a) the HYBRID dataset with 12 classes, including single- and hybrid-symptom images; and (b) the SINGLE dataset with 11 single-symptom classes. Diagonal cells indicate correct predictions, whereas off-diagonal cells indicate misclassifications; darker blue represents a higher prediction frequency.

remained the best-performing model overall, achieving a test precision of 84.40%, recall of 88.80%, F1-score of 86.00%, and accuracy of 86.00%.

Figure 10 presents a dumbbell plot comparing the AUC values of the 11 shared classes in the SINGLE and HYBRID datasets. Red rust had the lowest AUC in both datasets and was the class most adversely affected by the introduction of hybrid images. When red rust occurs as a secondary symptom, it may hinder the models' ability to distinguish the primary class. Moreover, red rust has a variable visual pattern and is distributed irregularly across the leaf surface; consequently, even models trained on the SINGLE dataset had difficulty classifying it. In this context, red rust may act as visual

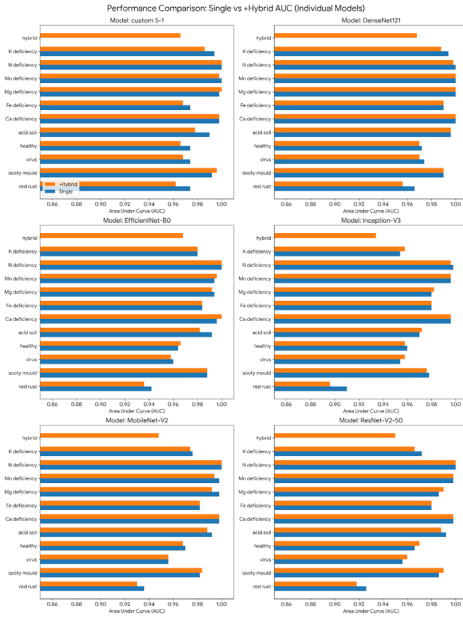


Fig. 9. Average AUC values for the six CNN models on the HYBRID dataset.

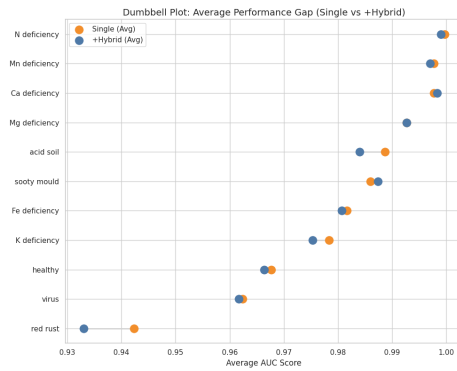


Fig. 10. AUC values for the 11 shared classes in the SINGLE and HYBRID datasets.

noise. By contrast, conditions with more consistent visual patterns, such as nitrogen deficiency, achieved near-perfect classification in the experiments.

B. Discussion

This study examined the effect of secondary disease symptoms on the performance of six CNN models. The models were evaluated using a dataset containing only single-symptom images and a second dataset that added images containing both potassium deficiency and red rust. The inclusion of hybrid images reduced classification performance, particularly for classes associated with the overlapping symptoms, and resulted in decreases in more than 40% of the class-level AUC values. Red rust was especially difficult to classify because of its variable appearance and irregular distribution across the leaf surface. Although excluding ambiguous images may improve performance in controlled experiments, such images are likely to occur in practice. Future work should therefore investigate

preprocessing, multilabel classification, or other methods that can explicitly account for co-occurring symptoms and improve diagnostic reliability under field conditions.

ACKNOWLEDGMENT

The authors thank the Malaysian Pepper Board for providing resources for data collection and image verification. Dr. Khew Choy Yuen and Dr. Lau Ee Tiing assisted with image labelling, and Associate Professor Dr. Hwang Siaw San assisted with data collection. The authors also acknowledge the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, for funding the publication fee. Figures 4–10 were generated or edited using Google Gemini.

REFERENCES

- [1] T. D. Salka, M. B. Hanafi, S. M. S. A. A. Rahman, D. B. M. Zulperi, and Z. Omar, "Plant leaf disease detection and classification using convolution neural networks model: a review," *Artificial Intelligence Review*, vol. 58, no. 10, p. 322, Jul 2025. [Online]. Available: <https://doi.org/10.1007/s10462-025-11234-6>
- [2] J. Chen, W. Chen, A. Zeb, S. Yang, and D. Zhang, "Lightweight inception networks for the recognition and detection of rice plant diseases," *IEEE Sensors Journal*, vol. 22, no. 14, pp. 14 628–14 638, 2022.
- [3] S. T. Y. Ramadan, T. Sakib, M. M. Ul Haque, N. Sharmin, and M. M. Rahman, "Generative adversarial network-based augmented rice leaf disease detection using deep learning," in *Proceedings of 2022 25th International Conference on Computer and Information Technology, ICCIT 2022*, 2022, pp. 976–981.
- [4] P. Tharani Pavithra and B. Baranidharan, "Resource prudent cnn models for disease identification of rice crops," in *Proceedings of the 1st IEEE International Conference on Networking and Communications 2023, ICNWC 2023*, 2023.
- [5] R. K. Vamsee Kongara, V. Siva Charan Somasila, N. Revanth, and R. D. Polagani, "Classification and comparison study of rice plant diseases using pre-trained cnn models," in *5th International Conference on Inventive Computation Technologies, ICICT 2022 - Proceedings*, 2022, pp. 140–145.
- [6] O. C. Hui Chen, C. H. Bong, and N. K. Lee, "Benchmarking cnn models for black pepper diseases and malnutrition prediction," in *2023 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, 2023, pp. 146–151.
- [7] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.
- [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 (ICML)*, 2019, pp. 6105–6114.
- [10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 630–645.
- [12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.
- [13] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320396001422>