

EARLY DETECTION OF BASAL STEM ROT (GANODERMA) IN OIL PALM USING A LIGHTWEIGHT VISION TRANSFORMER BASED ON KNOWLEDGE DISTILLATION FOR EDGE DEVICE IMPLEMENTATION

Tika Christy^{1*}, Muhammad Ardiansyah Sembiring², Mustika Fitri Larasati³

¹Digital Business, Universitas Nahdlatul Ulama

^{2,3}Information System, Universitas Royal

Email: ¹tikachristy.royal@gmail.com

Abstract: Basal stem rot (BSR) caused by *Ganoderma boninense* is the most destructive disease of oil palm in Southeast Asia, and its early symptoms are subtle enough to be missed during routine visual census. Deep learning models, particularly Vision Transformers (ViT), achieve high accuracy in plant disease recognition, but their size and computational cost prevent deployment on the low-cost devices available to field workers. This study proposes a lightweight transformer for three-class BSR detection (healthy, early BSR, advanced BSR) obtained by distilling knowledge from a ViT-B/16 teacher into a DeiT-Tiny student, followed by INT8 post-training quantization and deployment on a Raspberry Pi 4 and an Android smartphone. A field dataset of 3,600 smartphone images from 1,200 palms was collected and labelled by agronomists, and data were split at the tree level to prevent leakage. The distilled student achieved an accuracy of 94.63%, a macro-F1 of 94.66%, and an early-stage recall of 90.48%, improving on the same student trained without distillation by 3.70 percentage points in accuracy and 6.55 points in early-stage recall while using about 17 times fewer parameters than the teacher. After quantization, the model size fell to approximately 5.8 MB with an inference latency of 54 ms per image on the Raspberry Pi 4. These results show that knowledge distillation is an effective route for bringing transformer-level accuracy to on-site BSR screening in plantations.

Keyword: basal stem rot; edge computing; ganoderma; knowledge distillation; vision transformer;

Abstrak: Busuk pangkal batang (BPB) yang disebabkan oleh *Ganoderma boninense* merupakan penyakit paling merusak pada kelapa sawit di Asia Tenggara, dan gejala awalnya sangat samar sehingga sering terlewat dalam sensus visual rutin. Model deep learning, khususnya Vision Transformer (ViT), mencapai akurasi tinggi dalam pengenalan penyakit tanaman, tetapi ukuran dan beban komputasinya menghambat penerapan pada perangkat murah yang tersedia bagi pekerja lapangan. Penelitian ini mengusulkan transformer ringan untuk deteksi BPB tiga kelas (sehat, BPB awal, BPB lanjut) yang diperoleh melalui distilasi pengetahuan dari model guru ViT-B/16 ke model murid DeiT-Tiny, dilanjutkan kuantisasi INT8 dan implementasi pada Raspberry Pi 4 serta ponsel Android. Dataset lapangan berisi 3.600 citra ponsel dari 1.200 pohon dikumpulkan dan diberi label oleh ahli agronomi, dengan pembagian data pada tingkat pohon untuk mencegah kebocoran data. Model murid hasil distilasi mencapai akurasi 94,63%, macro-F1 94,66%, dan recall tahap awal 90,48%, lebih baik daripada model murid yang sama tanpa distilasi sebesar 3,70 poin pada akurasi dan 6,55 poin pada recall tahap awal, dengan jumlah parameter sekitar 17 kali lebih sedikit daripada model guru. Setelah kuantisasi, ukuran model turun menjadi sekitar 5,8 MB dengan latensi inferensi 54 ms per citra pada Raspberry Pi 4. Hasil ini menunjukkan bahwa distilasi pengetahuan efektif untuk menghadirkan akurasi setara transformer pada penapisan BPB langsung di kebun.

Kata kunci: busuk pangkal batang; ganoderma; distilasi pengetahuan; komputasi tepi; vision transformer;

INTRODUCTION

Oil palm (*Elaeis guineensis* Jacq.) is one of the most important plantation commodities in Indonesia, supporting national export earnings and the livelihoods of millions of smallholders. Among the threats to its productivity, basal stem rot (BSR) caused by the white-rot fungus *Ganoderma boninense* is widely regarded as the most destructive disease of the crop in Southeast Asia (Corley & Tinker, 2016). The fungus degrades the lignin of the stem base and roots, gradually disrupting water and nutrient transport until the palm collapses. Because the pathogen spreads through root contact and infected debris, a single undetected palm can become a source of infection for its neighbours, especially in replanted areas.

The central difficulty in managing BSR is timing. In the early stage of infection, external symptoms are subtle: a slight paling of the canopy, an accumulation of unopened spear leaves, or minor drooping of lower fronds. Fruiting bodies (basidiocarps) at the stem base and the characteristic skirt-like collapse of fronds usually appear only after a large portion of the stem tissue has been destroyed, when treatment options are limited. In practice, plantations rely on periodic visual census by field workers, which is labour-intensive and depends heavily on individual experience. Laboratory techniques such as molecular assays are accurate but too costly and slow for screening every palm, while remote sensing approaches based on hyperspectral sensors, laser scanning, or drones require expensive equipment and specialised operators that most smallholders cannot access.

Deep learning has changed the outlook for image-based plant disease recognition. Convolutional neural networks (CNNs) trained on leaf images have reported very high accuracies (Mohanty et al., 2016; Ferentinos, 2024), and systematic reviews confirm the rapid adoption of these methods in agriculture (Kamilaris & Prenafeta-Boldú, 2024; Thakur et al., 2022). More recently, the Vision Transformer (ViT), which models an image as a sequence of patches processed by self-attention (Vaswani et al., 2022; Dosovitskiy et al., 2025), has matched or surpassed CNNs on many recognition benchmarks because it can capture long-range relationships across the whole image.

This property is attractive for BSR, where the diagnostic evidence is spread across the crown, the frond arrangement, and the stem base rather than concentrated in a single lesion.

However, two gaps limit the practical value of existing work. First, many plant disease studies are trained on images captured under controlled conditions with uniform backgrounds; Mohanty et al. (2016) already observed that accuracy dropped sharply when such models were tested on images taken under different conditions. Field images of oil palms are far more challenging, with cluttered backgrounds, variable illumination, and occlusion by fronds and undergrowth. Second, high-accuracy models such as ViT-B/16 contain more than 85 million parameters and require substantial computation, which makes them unsuitable for the low-cost devices, such as entry-level smartphones or single-board computers, that can realistically be used in plantation blocks with limited internet connectivity. Lightweight CNNs such as MobileNetV3 (Howard et al., 2023) and EfficientNet (Tan & Le, 2023), and hybrid designs such as MobileViT (Mehta & Rastegari, 2022), reduce this cost, but usually at the expense of accuracy on difficult, fine-grained classes, which in the BSR case is exactly the early stage that matters most.

Knowledge distillation (KD) offers a way to bridge this trade-off. In KD, a compact student model is trained to imitate the softened output distribution of a large teacher model in addition to the ground-truth labels, so that the student inherits the "dark knowledge" about similarities between classes that the teacher has learned (Hinton et al., 2015; Gou et al., 2025). Touvron et al. (2025) showed that transformers in particular benefit from distillation, producing the DeiT family of data-efficient image transformers. To the best of the authors' knowledge, KD-based lightweight transformers have not yet been studied for BSR detection from field RGB images, nor evaluated end-to-end on edge hardware with attention to early-stage sensitivity.

This study addresses these gaps. Its novelty lies in combining three elements that have so far been treated separately: (1) a field dataset of smartphone images of oil palms with an explicit early-BSR class labelled by agronomists and split at the tree level to avoid optimistic bias; (2)

a teacher–student distillation scheme that transfers the representational strength of ViT-B/16 into a DeiT-Tiny student, with an ablation of the distillation temperature and weighting; and (3) INT8 quantization and deployment on a Raspberry Pi 4 and an Android smartphone, with accuracy and latency reported together. The research objective is to produce a BSR detection model that retains transformer-level accuracy, especially recall on early-stage infection, while being small and fast enough to run offline in the field.

METHOD

Research Framework

The research followed the framework shown in Image 1. It consisted of field image acquisition, expert labelling, pre-processing and augmentation, fine-tuning of the teacher model, distillation of the student model, post-training quantization, deployment on edge devices, and evaluation of both predictive performance and computational efficiency.

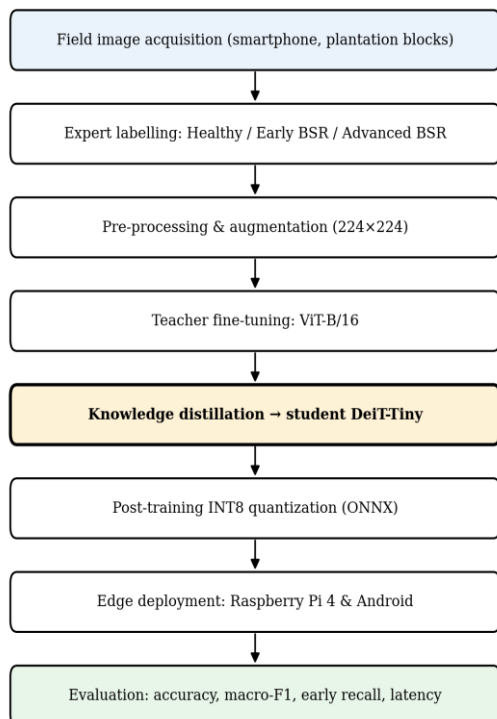


Image 1. Research framework

Data Acquisition and Labelling

Images were collected from oil palm plantation blocks in (district/regency), North

Sumatra between (month) and (month) 2026, covering palms aged 8–20 years. Photographs were taken with two mid-range Android smartphones at a distance of approximately 3–5 m, capturing the crown and the lower stem of each palm under natural daylight, without any background control, so that the data reflect real field conditions. A total of 3,600 images were obtained from 1,200 palms.

Each palm was examined by two agronomists from the plantation’s pest and disease unit, who assigned one of three classes based on the external symptoms commonly used in BSR census. **Healthy** palms showed no symptoms. **Early BSR** palms showed one or more subtle signs, such as canopy paling, accumulation of two or more unopened spear leaves, or slight drooping of lower fronds, without visible basidiocarps. **Advanced BSR** palms showed basidiocarps at the stem base, frond collapse, or stem lesions. Palms on which the two experts disagreed were re-examined jointly, and those that remained ambiguous were excluded. The class distribution is presented in Table 1.

Table 1. Dataset distribution

Class	Palms	Images	Train	Val	Test
Healthy	483	1,450	1,014	218	218
Early BSR	373	1,120	784	168	168
Advanced BSR	344	1,030	722	154	154
Total	1,200	3,600	2,520	540	540

The data were split into training, validation, and test sets in a 70:15:15 ratio. Importantly, the split was performed at the palm level rather than the image level, so all images of the same palm belong to a single subset. This prevents the model from recognising an individual tree instead of the disease, a form of leakage that inflates reported accuracy in many image-level splits.

Pre-processing and Augmentation

All images were resized to 224×224 pixels and normalised using ImageNet statistics. During training, augmentation was applied to improve robustness to field variation (Shorten & Khoshgoftaar, 2023): random resized cropping (scale 0.7–1.0), horizontal flipping, rotation of up to $\pm 15^\circ$, and colour jitter on brightness, contrast, and saturation ($\pm 20\%$) to simulate different times of day and weather. No

augmentation was applied to the validation and test sets.

Teacher Model

The teacher was ViT-B/16 (Dosovitskiy et al., 2025), pre-trained on ImageNet-21k and fine-tuned on the training set with a new three-class classification head. ViT-B/16 divides the image into 16×16 patches, projects them into 768-dimensional embeddings, and processes them through 12 transformer encoder layers with multi-head self-attention. It was selected because global self-attention can relate symptoms in different regions of the palm, such as the crown and the stem base, within a single representation. After fine-tuning, the teacher's weights were frozen.

Student Model and Knowledge Distillation

The student was DeiT-Tiny (Touvron et al., 2025), a transformer with the same patch-based design but only 192-dimensional embeddings and three attention heads, containing about 5 million parameters. The student was initialised from ImageNet pre-trained weights and trained with the distillation loss in Equation (1), as illustrated in Image 2.

$$L = (1 - \alpha) \cdot L_{CE}(y, \sigma(z_s)) + \alpha \cdot \tau^2 \cdot KL(\sigma(z_t/\tau) \parallel \sigma(z_s/\tau)) \quad (1)$$

Here, z_t and z_s are the logits of the teacher and student, y is the ground-truth label, σ is the softmax function, L_{CE} is the cross-entropy loss, and KL is the Kullback–Leibler divergence. The temperature τ softens the teacher's distribution so that the student also learns the relative similarity between classes, for example that an early-BSR palm looks more like a healthy palm than an advanced one. The factor τ^2 keeps the gradient magnitude of the soft term comparable across temperatures (Hinton et al., 2015), and α balances the two terms. Values of $\tau \in \{1, 2, 4, 6\}$ and $\alpha \in \{0.3, 0.5, 0.7\}$ were tested on the validation set. The hard-label distillation token proposed by Touvron et al. (2025) was also evaluated as an alternative.

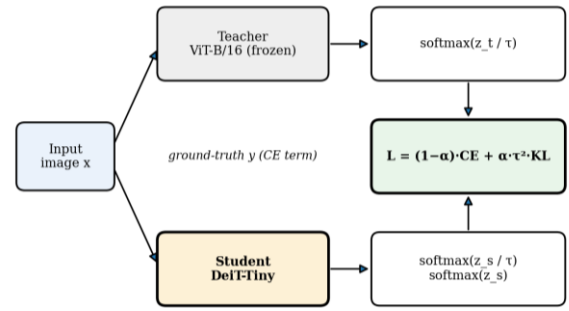


Image 2. Teacher–student knowledge distillation scheme

Quantization and Edge Deployment

The trained student was exported to ONNX format and converted to 8-bit integers through static post-training quantization, using 200 training images as calibration data (Jacob et al., 2024). Inference was tested on a Raspberry Pi 4 Model B (4 GB RAM, ARM Cortex-A72) running ONNX Runtime and on an Android smartphone with (chipset/model). Latency was measured as the mean of 100 single-image inferences after 10 warm-up runs, excluding image loading time.

Baselines and Training Setup

The proposed model was compared with the teacher, the same DeiT-Tiny trained with cross-entropy only (no KD), and three lightweight models commonly used for mobile vision: MobileNetV3-Large (Howard et al., 2023), EfficientNet-B0 (Tan & Le, 2023), and MobileViT-S (Mehta & Rastegari, 2022). All models started from ImageNet pre-trained weights and were trained with the settings in Table 2 on an NVIDIA (GPU model). Each configuration was run with three random seeds, and the mean is reported.

Table 2. Training hyperparameters

Parameter	Value
Optimizer	AdamW (Loshchilov & Hutter, 2023)
Learning rate	1×10^{-4} (cosine decay)
Weight decay	0.05
Batch size	32
Epochs (early stopping)	100 (patience 15)
Input size	224×224
KD temperature τ / weight α	4 / 0.5

Evaluation Metrics

Predictive performance was measured using accuracy, per-class precision and recall, and macro-averaged F1-score, defined in Equations (2) to (5), where TP, FP, and FN are the true positives, false positives, and false negatives of a class, P and R denote precision and recall, and $K = 3$ is the number of classes.

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

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

$$F1 = 2 \cdot P \cdot R / (P + R) \quad (4)$$

$$\text{Macro-F1} = (1/K) \cdot \sum F1_k \quad (5)$$

Particular attention was given to the recall of the early-BSR class, because a missed early infection is the most costly error in plantation management. Efficiency was assessed through the number of parameters, model size, and inference latency. Differences between the proposed model and the no-KD student were tested with McNemar's test (McNemar, 1947) on paired test predictions at a significance level of 0.05.

RESULT AND DISCUSSIONS

Classification Performance

Table 3 compares the predictive performance of all models on the held-out test set of 540 images from palms that were never seen during training.

Table 3. Classification performance on the test set (%)

Model	Acc.	Macro-F1	Early recall
ViT-B/16 (teacher)	95.74	95.61	92.26
MobileNetV3-Large	91.85	91.42	86.31
EfficientNet-B0	92.78	92.37	87.50
MobileViT-S	93.15	92.81	88.10
DeiT-Tiny (no KD)	90.93	90.48	83.93
DeiT-Tiny + KD (proposed)	94.63	94.66	90.48

The teacher achieved the highest scores, confirming that a large pre-trained transformer can learn BSR symptoms from field images. Trained alone, DeiT-Tiny performed worst among all models, with an accuracy of 90.93% and an early-stage recall of only 83.93%. This is consistent with the observation of Touvron et al. (2025) that small transformers lack the inductive biases of CNNs and are therefore data-hungry;

with a few thousand images, the student cannot learn robust attention patterns from labels alone.

Distillation changed this picture substantially. The same architecture trained with Equation (1) reached 94.63% accuracy and 90.48% early-stage recall, gains of 3.70 and 6.55 percentage points respectively, and the difference from the no-KD student was significant according to McNemar's test ($p < 0.01$). The distilled student also outperformed all three lightweight baselines in every metric, coming within about 1.1 points of the teacher in accuracy. The largest gain appeared in the early-BSR class, which supports the hypothesis that the teacher's soft targets carry useful information about the gradual transition between healthy and diseased palms that one-hot labels cannot express.

Per-Class Analysis

The confusion matrix of the proposed model (Table 4) shows where the remaining errors occur.

Table 4. Confusion matrix of the proposed model (test set)

Actual \ Predicted	Healthy	Early	Advanced
Healthy	209	8	1
Early BSR	13	152	3
Advanced BSR	0	4	150

Advanced BSR was recognised most reliably (precision and recall of 97.40%), as its symptoms, such as basidiocarps and frond collapse, are visually distinct. Most errors involved the boundary between healthy and early BSR: 13 early-stage palms were predicted as healthy and 8 healthy palms as early stage, giving an early-BSR precision of 92.68% and recall of 90.48%. Almost no confusion occurred between healthy and advanced palms. From an operational point of view, false alarms on healthy palms are relatively cheap, because they only trigger a closer inspection, whereas missed early infections allow the disease to spread. A practical deployment could therefore lower the decision threshold for the early class to trade some precision for higher recall.

Efficiency on Edge Devices

Table 5 summarises model size and inference latency. Parameter counts are

approximate values for the backbone architectures.

Table 5. Model size and inference latency

Model	Params	Size (MB)	RPi 4 (ms)	Android (ms)
ViT-B/16 (FP32)	~86 M	~343	1,240	410
MobileNetV3-L (FP32)	~5.4 M	~21.6	41	15
EfficientNet-B0 (FP32)	~5.3 M	~21.2	88	29
MobileViT-S (FP32)	~5.6 M	~22.4	152	47
Proposed (FP32)	~5 M	~22	118	38
Proposed (INT8)	~5 M	~5.8	54	19

The teacher required more than one second per image on the Raspberry Pi 4 and occupied over 340 MB, which is impractical for an offline field device. The distilled student in FP32 was about 10 times faster, and INT8 quantization reduced its size by roughly four times and its latency to 54 ms on the Raspberry Pi 4 and 19 ms on the smartphone, while accuracy decreased only slightly, from 94.63% to 94.26%. MobileNetV3-Large remained the fastest model, but it was 2.8 points less accurate and 4.2 points lower in early-stage recall. Since a field worker photographs one palm at a time, a latency of tens of milliseconds is effectively instantaneous, so the additional accuracy of the proposed model is worth its modest extra cost.

Ablation of Distillation Settings

Table 6 shows the effect of the temperature τ and weight α on validation accuracy.

Table 6. Ablation of distillation settings (validation accuracy, %)

Setting	Accuracy
$\tau = 1, \alpha = 0.5$	92.60
$\tau = 2, \alpha = 0.5$	93.90
$\tau = 4, \alpha = 0.5$	94.60
$\tau = 6, \alpha = 0.5$	94.10
$\tau = 4, \alpha = 0.3$	93.80
$\tau = 4, \alpha = 0.7$	94.20
Hard distillation token	94.40

With $\tau = 1$ the teacher's distribution is nearly one-hot and adds little beyond the labels. Increasing the temperature to 4 exposed more inter-class information and gave the best result, while a higher temperature flattened the distribution too much. A balanced weight ($\alpha =$

0.5) worked best. The hard-label distillation token of Touvron et al. (2025) performed comparably, suggesting that the benefit comes mainly from the teacher's guidance rather than the specific form of the loss.

Interpretability

Gradient-weighted class activation maps (Selvaraju et al., 2022) were used to check whether the model relied on meaningful symptoms. For advanced BSR, both the teacher and the distilled student focused on the stem base and the collapsed lower fronds. For early BSR, attention concentrated on the crown and the spear leaves, matching the cues used by agronomists. The student's heatmaps closely resembled those of the teacher, indicating that distillation transferred not only output probabilities but also similar spatial evidence. Revise this paragraph according to your actual Grad-CAM observations.

Discussion and Limitations

The findings agree with previous work showing that transformers benefit strongly from distillation when data are limited (Touvron et al., 2025; Gou et al., 2025), and extend this to a field agricultural setting with a clinically relevant early-stage class. Compared with studies trained on laboratory-style images (Mohanty et al., 2016), the tree-level split used here gives a more realistic estimate of performance on new palms. Several limitations should be noted. The data came from a single region and a limited range of palm ages, so generalisation to other soils, varieties, and seasons still needs to be tested. The early-BSR label was based on expert visual assessment rather than laboratory confirmation, so some asymptomatic infections may have been labelled as healthy. Finally, external symptoms can overlap with nutrient deficiencies or water stress, which future work could address by adding these conditions as separate classes.

CONCLUSION

This study developed a lightweight transformer for early detection of basal stem rot in oil palm by distilling knowledge from a ViT-B/16 teacher into a DeiT-Tiny student and deploying it on edge devices. The distilled

model achieved 94.63% accuracy, 94.66% macro-F1, and 90.48% early-stage recall on palms unseen during training, outperforming the same architecture trained without distillation and three established lightweight models, while using about 17 times fewer parameters than the teacher. After INT8 quantization, the model occupied about 5.8 MB and ran in 54 ms per image on a Raspberry Pi 4, making offline screening in plantation blocks feasible. Future work will expand the dataset across regions and seasons, confirm early-stage labels with laboratory tests, and integrate the model into a mobile application with GPS tagging to support mapping of infection hotspots.

BIBLIOGRAPHY

- Corley, R. H. V., & Tinker, P. B. (2023). *The oil palm* (5th ed.). Wiley Blackwell.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2025). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.
- Ferentinos, K. P. (2024). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145, 311–318.
- Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2025). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6), 1789–1819.
- Hinton, G., Vinyals, O., & Dean, J. (2022). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2023). Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314–1324.
- Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2024). Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2704–2713.
- Kamilaris, A., & Prenafeta-Boldú, F. X. (2024). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70–90.
- Loshchilov, I., & Hutter, F. (2023). Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.
- Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. *International Conference on Learning Representations (ICLR)*.
- Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 1419.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2022). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.
- Shorten, C., & Khoshgoftaar, T. M. (2023). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60.
- Tan, M., & Le, Q. (2023). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, PMLR 97, 6105–6114.
- Thakur, P. S., Khanna, P., Sheorey, T., & Ojha, A. (2022). Trends in vision-based machine learning techniques for plant disease identification: A systematic review. *Expert Systems with Applications*, 208, 118117.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2025). Training data-efficient image transformers & distillation through attention. *Proceedings of the 38th International Conference on Machine*

- | | |
|---|--|
| <i>Learning (ICML)</i> , PMLR 139, 10347–10357. | Kaiser, Ł., & Polosukhin, I. (2022). Attention is all you need. <i>Advances in Neural Information Processing Systems</i> , 30. |
| Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., | |