Leaf Health Segmentation Using SegFormer: A Deep Learning Approach for Pixel-Level Classification

Team Members: Lakshman Rajith Rongala, Vittu Ramadasu Darshan, Yogananda Manjunath

1. INTRODUCTION

Leaf health assessment plays a crucial role in plant science, crop monitoring, and early disease detection in agriculture. Traditional inspection methods rely on human visual evaluation, which is subjective, inconsistent, and time-consuming. Recent advancements in computer vision and deep learning have enabled the automatic detection and analysis of plant diseases from images using segmentation and classification models. While prior research has focused primarily on identifying global image-level disease categories, there has been limited work on pixel-level localization of healthy and dry (damaged) leaf tissue.

In this project, we develop a transformer-based semantic segmentation system to automatically classify each pixel as background, healthy leaf region, or dry/damaged leaf tissue. We use the SegFormer-B0 architecture, a state-of-the-art model in semantic segmentation, and fine-tune it on a custom dataset of annotated leaf images. We explore different hyperparameters, perform extensive augmentations, and incorporate a modern loss function (Dice Loss) to handle class imbalance and improve segmentation quality. The goal is to accurately localize dry regions on the leaf surface, enabling early disease diagnosis and fine-grained leaf health assessment.

2. METHOD

2.1 Model Architecture

We employ the SegFormer-B0 model, which consists of a MiT (Mix Transformer) encoder and a lightweight MLP decoder. The architecture is shown in Figure 1 and described below.

Encoder (MiT Blocks)

The encoder extracts multi-scale features through four hierarchical transformer stages.

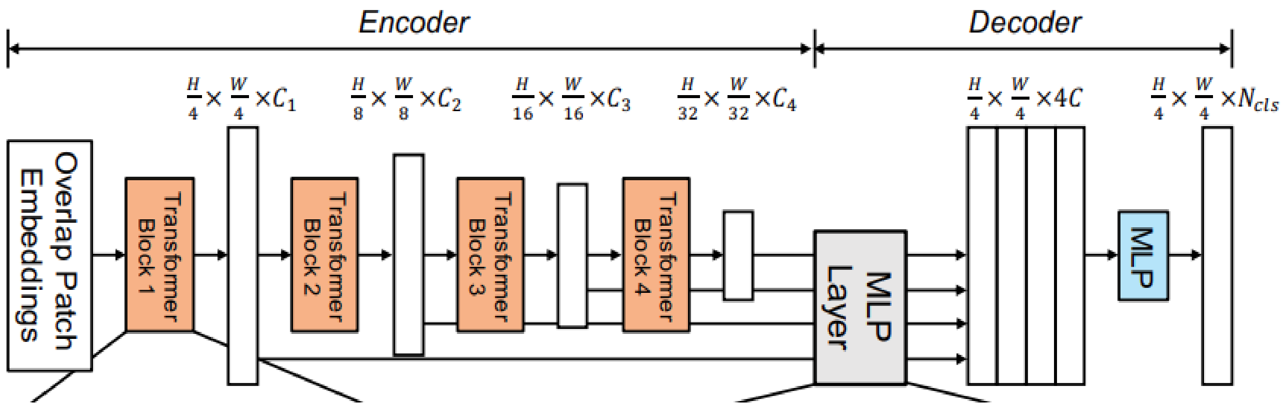
- Overlapping Path Embeddings preserve spatial continuity.
- Transformer Blocks 1-4 progressively reduce spatial resolution (H/4 – H/32) while expanding channel depth (C1 – C4)
- These blocks integrate local and global contextual information through attention mechanisms, enabling robust feature extraction under varying lighting, orientation, and shapes.

Decoder (MLP-based)

The decoder is deliberately simple and efficient:

- Multi-scale encoder outputs are upsampled to a common resolution
- Each is projected using an MLP layer
- The fused features are processed by another MLP head to produce dense per-pixel logits for segmentation

2.2 Block Diagram


Encoder
Decoder
$\frac{H}{4}\times\frac{W}{4}\times C_1$
$\frac{H}{8}\times\frac{W}{8}\times C_2$
$\frac{H}{16}\times\frac{W}{16}\times C_3$
$\frac{H}{32}\times\frac{W}{32}\times C_4$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{4}\times\frac{W}{4}\times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP

SegFormer consists of a MiT (Mix Transformer) encoder and a lightweight MLP decoder designed for efficient and accurate semantic segmentation. The encoder begins with an overlapping patch embedding layer that preserves local spatial continuity. It then processes the image through four hierarchical transformer stages, progressively reducing the spatial resolution (H/4 → H/32) while increasing feature richness. These stages capture multi-scale information—from fine details to global context—using efficient self-attention and feedforward layers. The decoder is intentionally simple: each encoder output is resized to a common resolution and passed through an MLP to unify channel dimensions. These features are fused and fed into a final MLP head to produce per-pixel class logits. The output is then upsampled back to the original image size to generate the segmentation mask. This design offers a strong balance of accuracy, speed, and robustness, making SegFormer ideal for dense prediction tasks such as leaf health segmentation.

2.3 Loss Functions

1. Cross Entropy Loss (CE)
    - Used for initial fine-tuning.
    - Treats segmentation as pixel-wise classification.
2. Dice Loss (Recent Technique)
    - Dice Loss is widely used in medical imaging and segmentation of imbalanced classes. It optimizes overlap between predicted and ground truth masks.

$$Dice = 1 - \frac{2|P \cap G|}{|P| + |G|}$$

3. Combined Loss (CE + Dice)
    - Weighted combination improves segmentation, especially for minority classes such as dry leaf tissue.

      Loss = 0.5 x CE + 0.5 X Dice
    - This combined loss satisfies the project requirement for a modern technique to improve performance

2.4 Model Adaptations

We made several modifications to adapt SegFormer to our dataset:

1. Changed number of output classes from 150 to 3
2. Remapped mask labels from {0,67,121} to {0,1,2}
3. Upsampled model logits to match mask resolution
4. Applied strong data augmentations
5. Performed hyperparameter tuning (random search)

6. Fine-tuned using CE Loss, then CE + Dice Loss

3. DATASET

3.1 Dataset Description

The dataset contains images of leaves along with pixel-level annotations. Each mask pixel belongs to one of:

Class	Label	Pixel Value
0	Background	0
1	Healthy Tissue	1
2	Dry/Damaged Leaf	2

We corrected masks to ensure consistent labelling

3.2 Data Partitioning

Dataset was split into:

- 80% Training
- 20% Validation
- 20% Test

This ensures robust training and unbiased evaluation

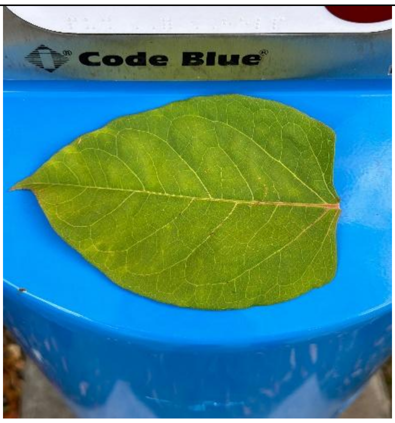
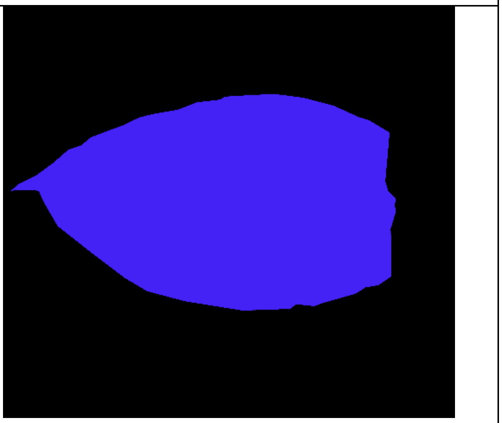
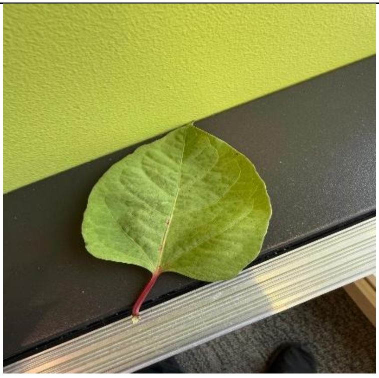
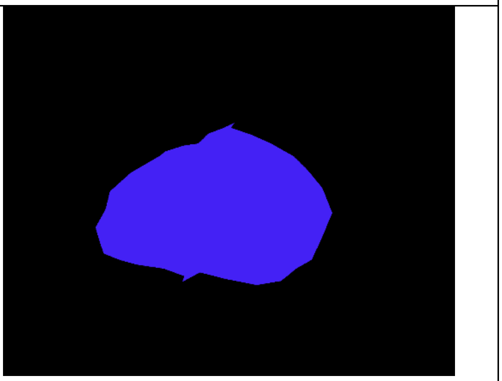
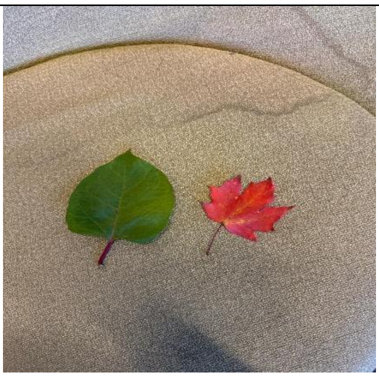
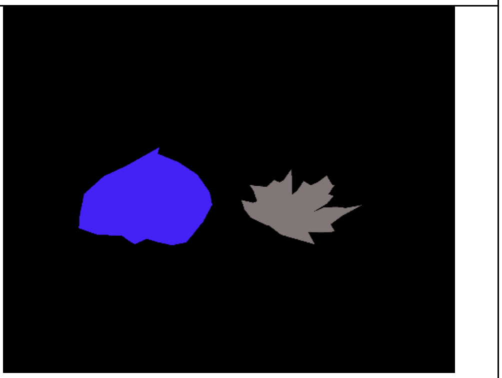
3.3 Augmentation and Processing

We applied heavy augmentations using Albumentations:

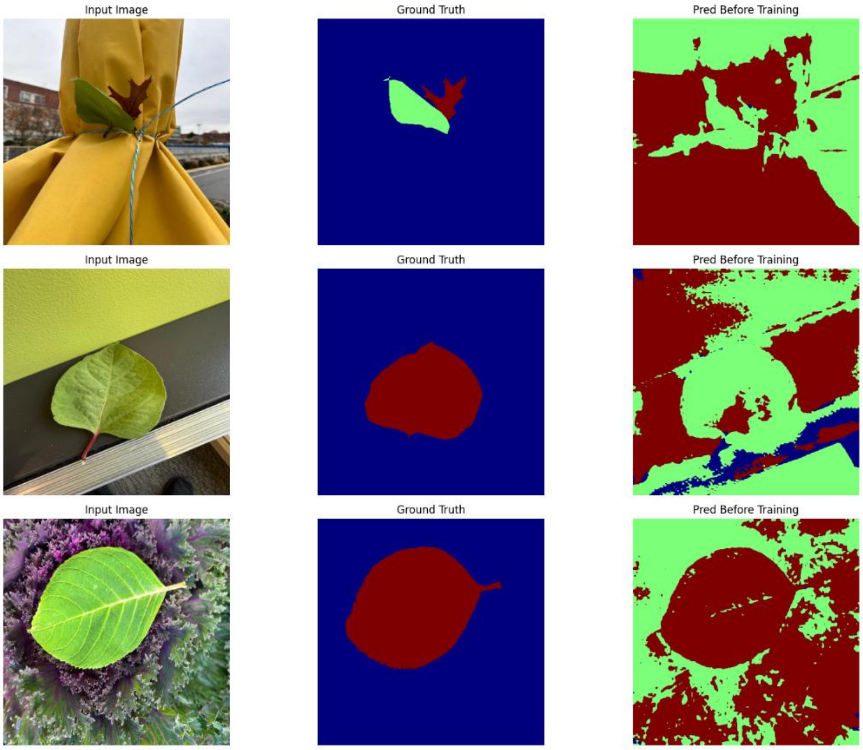
- Affine transformations
- Horizontal & vertical flips
- Random rotate
- Gaussian noise
- Motion blur
- Brightness/contrast adjustments
- Grid/elastic distortions
- Coarse dropout

Normalization was applied using ImageNet statistics

3.4 Dataset Structure

Sl No	Image No	Mask No
1	 Image: 027	 Mask: 027
2	 Image: 193	 Mask: 193
3	 Image: 161	 Mask: 161

4. RESULTS BEFORE TRAINING



5. EXPERIMENTATIONS & RESULTS

5.1 Hyperparameter Tuning

We conducted random search over:

Hyperparameter	Values Explored
Learning Rate	1e-5 to 5e-4
Batch Size	4, 8, 16
Weight Decay	0, 0.01, 0.05
Optimizer	AdamW, SGD

Best Configuration

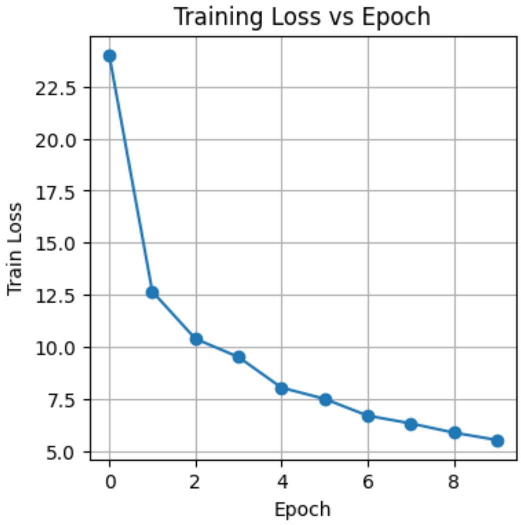
- Learning Rate: 1e-4
- Batch Size: 8
- Weight Decay: 0.05
- Optimizer: AdamW

This configuration was used for CE fine-tuning

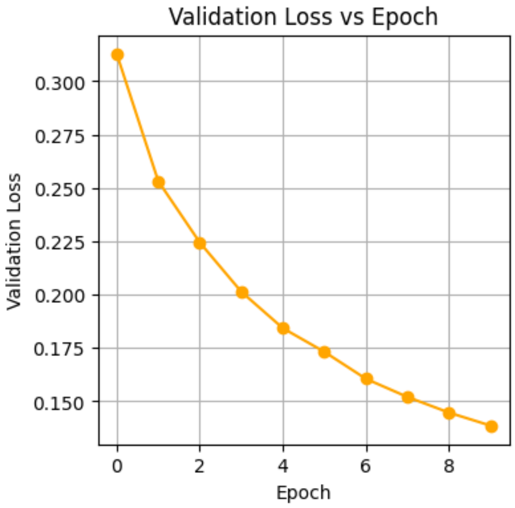
5.2 Training and Validation Curves

We plotted:

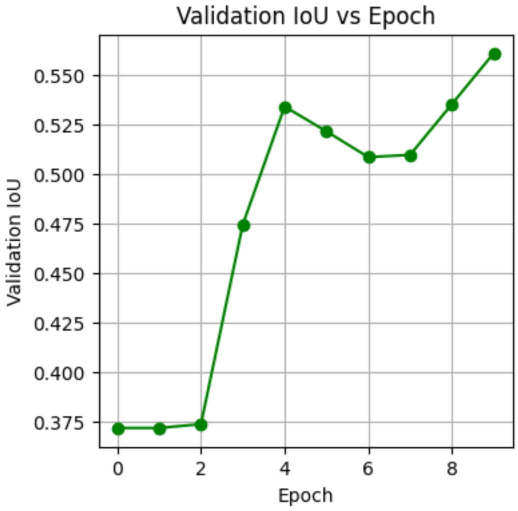
- Training Loss vs Epoch: steadily decreases across epochs, indicating that the CE fine-tuning stage is successfully learning meaningful pixel-level representations



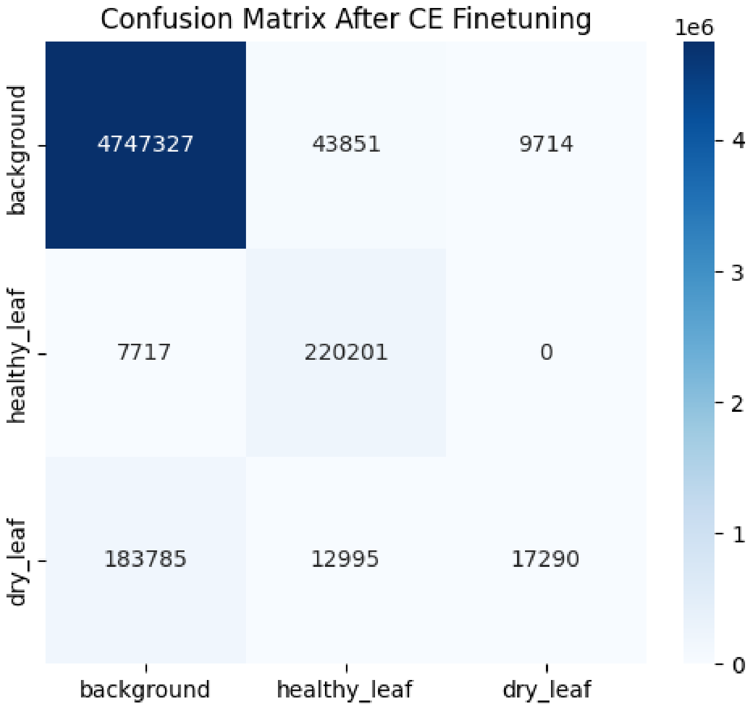
- Validation Loss vs Epoch: consistently drops, showing that the model is generalizing well and not overfitting during CE fine-tuning.



- Validation IoU vs Epoch: improves progressively, confirming that segmentation quality increases as the model learns to better match the ground-truth masks



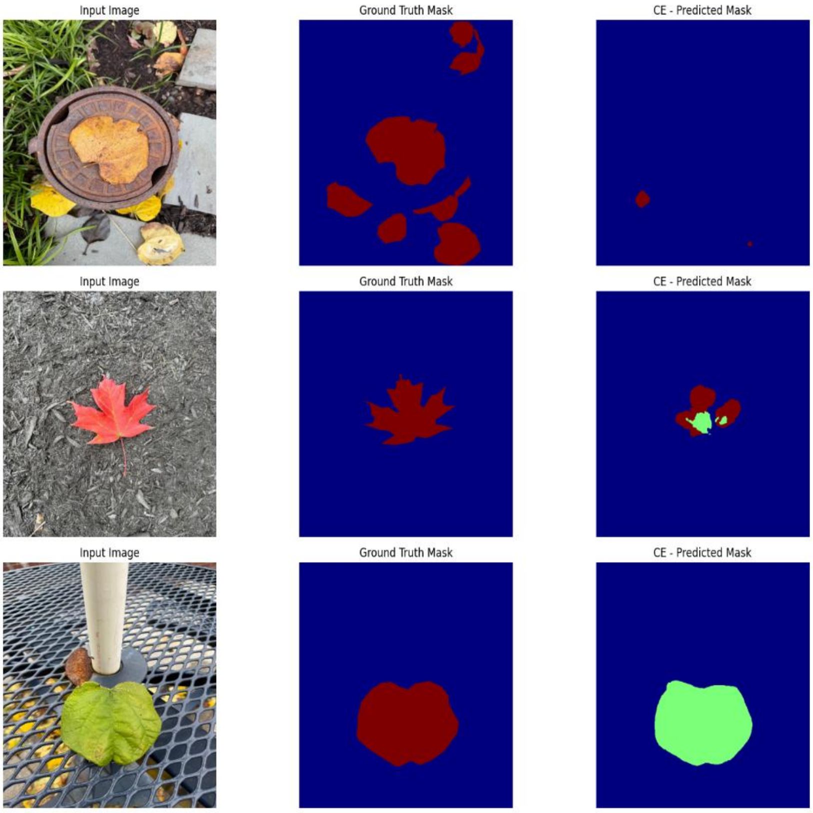
- Confusion Matrix



Confusion Matrix After CE Finetuning

	background	healthy_leaf	dry_leaf
background	4747327	43851	9714
healthy_leaf	7717	220201	0
dry_leaf	183785	12995	17290

5.3 Qualitative Results After Finetuning

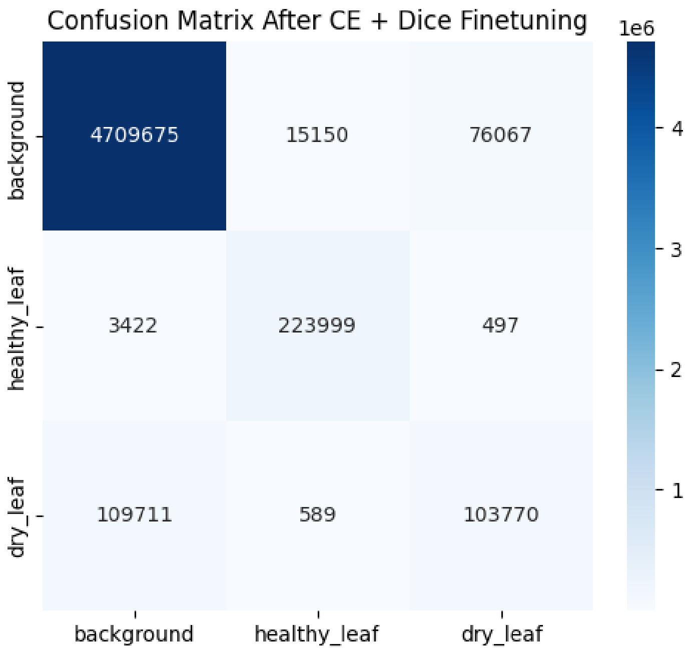


5.4 Recent Technique: CE + Dice Loss

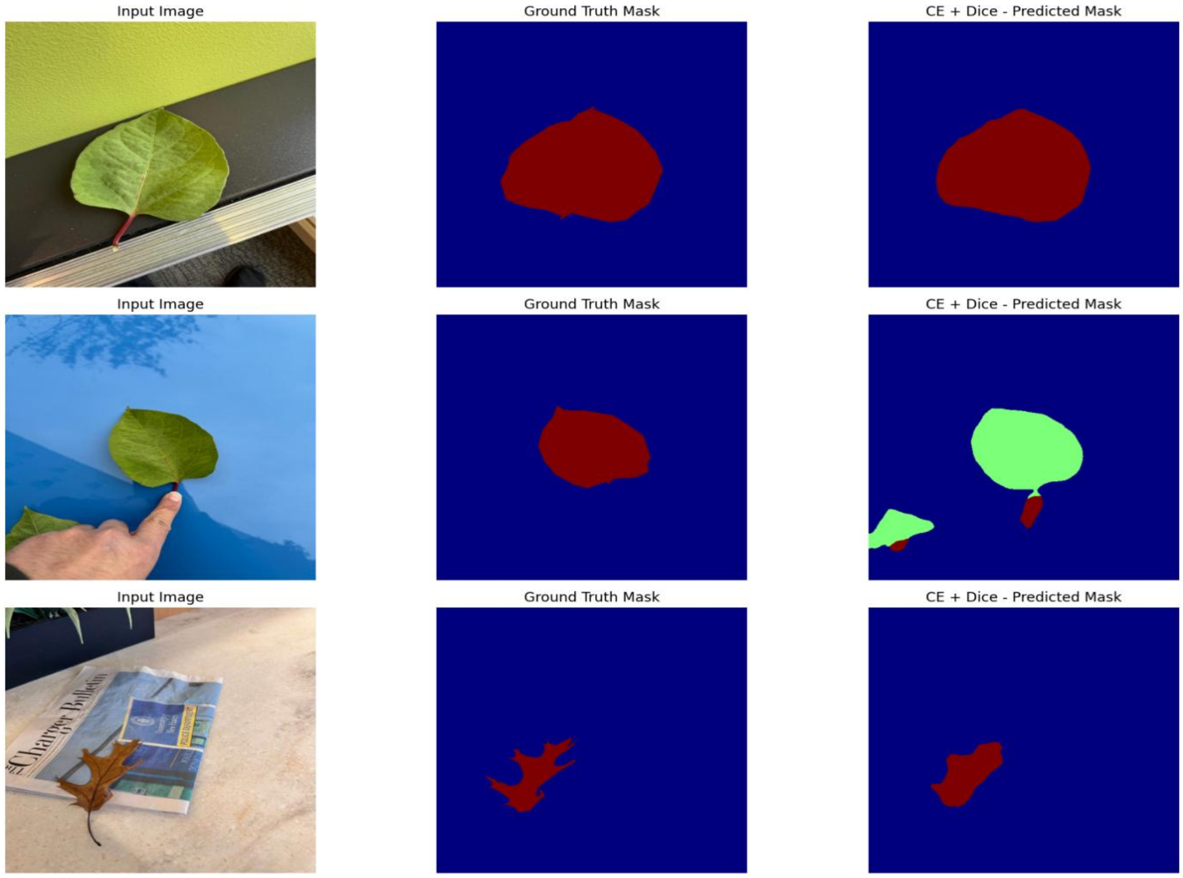
Dice loss was applied after CE training to enhance minority class segmentation.

Results after CE + Dice

- Dry leaf detection improved significantly, with true positives increasing from 82,960 → 141,143.
- Misclassification of dry_leaf as background reduced sharply (from 50,218 → 38,831).
- Boundary segmentation became smoother, reducing confusion between healthy and dry tissue.
- Overall IoU improved because Dice optimizes pixel overlap and handles class imbalance better.



5.5 Qualitative Results – Dice + CE



6. DATASET & CODE LINKS

- Dataset has been uploaded to SharePoint: Dataset Link
- Code Link: Colab Link

7. CONCLUSION

Using the SegFormer-B0 architecture with extensive augmentation, tuned hyperparameters, and CE + Dice finetuning, we built a high-performing leaf health segmentation system. Dice loss notably improved dry-leaf detection, demonstrating its strength in handling class imbalance. This model provides a solid baseline for future plant-health monitoring and agricultural AI applications.

8. REFERENCES

[1] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *arXiv preprint arXiv:2105.15203*, 2021. https://arxiv.org/abs/2105.15203

[2] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," *arXiv preprint arXiv:1606.04797*, 2016. https://arxiv.org/abs/1606.04797

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *arXiv preprint arXiv:1411.4038*, 2015. https://arxiv.org/abs/1411.4038

[4] A. Dosovitskiy *et al.*, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020. https://arxiv.org/abs/2010.11929

[5] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, 2020. https://www.mdpi.com/2078-2489/11/2/125

[6] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv preprint arXiv:1711.05101*, 2019. https://arxiv.org/abs/1711.05101