ATTENTION-GATED U-NET WITH TRANSFER LEARNING AND A MULTI-FACTOR SEVERITY INDEX FOR PLANT LEAF DISEASE SEGMENTATION

Ms Pallavi B. Patil

Student ,M.Tech CSE(Data Science),D.Y.Patil Agriculture and Technical University ,Talsande,Kolhapur,Maharashtra,India

pallavipatil2515@gmail.com

Mr Abhijit S. Shinde

Assistant Professor, D.Y.Patil Agriculture and Technical University, Talsande, Kolhapur, Maharashtra, India

director.sa@dyp-atu.edu.in

Abstract : Automated plant disease systems overwhelmingly report image-level classification, which tells a grower that a leaf is diseased but not where the lesions sit, how far they have spread, or whether the model can be trusted. This work trains three segmentation architectures on the 588-image Kaggle Leaf Disease Segmentation dataset under an identical protocol and reports what each addition actually buys. A vanilla U-Net reaches 0.586 IoU. Swapping the encoder for an ImageNet-pretrained EfficientNet-B0 raises this to 0.681, the single largest gain in the study. Adding attention gates to the skip connections moves IoU to 0.670 while lifting precision from 0.822 to 0.871, a trade that reduces false positive lesion pixels at the cost of some recall. On top of the segmentation output we define a Multi-Factor Severity Index that combines diseased area, lesion count from connected component analysis, boundary circularity, and the spatial dispersion of lesion centroids. The index correlates with ground truth severity at Pearson r = 0.875 and assigns the correct grade in 80.5% of validation images, with a mean absolute error of 3.86 percentage points in predicted diseased area. Corruption tests show graceful degradation under mild noise and blur and a sharper 24.6-point IoU drop under severe brightness shift. We also flag a methodological issue: prior work on this dataset reports 0.943 IoU after augmenting to 7056 images and then splitting 70:30, a procedure that places augmented copies of the same leaf in both partitions. Our lower numbers come from splitting the 588 originals first and augmenting only the training fold.

IndexTerms - semantic segmentation, attention gates, transfer learning, plant pathology, severity estimation, precision agriculture, EfficientNet

I. INTRODUCTION

Pathogens and pests remove a fifth to two fifths of the global yield of wheat, rice, maize, potato, and soybean every year [1]. The figure is not stable across regions or seasons, and the losses fall hardest on smallholder farms that have the least capacity to absorb them. Detection still depends largely on someone walking the field and looking at leaves. That person is expensive, slow to reach remote plots, and inconsistent: inter-rater agreement on visual severity scoring frequently falls below a Cohen kappa of 0.6, which means two trained assessors looking at the same lesion often disagree about how bad it is [2].

Convolutional networks changed what is possible here. Mohanty et al. [3] showed that a CNN could separate 26 diseases across 14 crops from PlantVillage images, and a large literature has since pushed classification accuracy above 95% on that benchmark [4]. The trouble is that classification answers only one question. A model that outputs the label early blight has said nothing about which part of the leaf is affected, and a grower deciding between removing three leaves and spraying an entire row needs exactly that information. Precision spraying, which is where the environmental and economic argument for this technology actually lives, requires a spatial map.

Semantic segmentation supplies the map. U-Net [5] remains the reference architecture: a contracting encoder builds semantic understanding, an expanding decoder recovers resolution, and skip connections carry fine spatial detail across the bottleneck. Skip connections are also the architecture's weak point. They pass every encoder feature to the decoder, including the large majority that describe healthy tissue, leaving the decoder to learn what to ignore. Oktay et al. [6] addressed this with attention gates, small learned

modules that weight each spatial location of a skip connection by its relevance before concatenation. Attention U-Net has since been applied widely in medical imaging; its behaviour on small agricultural datasets is less well characterised.

A second gap concerns severity. Even papers that do segment lesions typically stop at a ratio: diseased pixels over total pixels. This treats a leaf with one 20% lesion and a leaf with twenty scattered lesions summing to 20% as equivalent cases, which no plant pathologist would accept. The first is localised and can be pruned. The second suggests systemic infection or repeated inoculation and calls for a different intervention.

This paper makes four contributions. First, we run a controlled three-way comparison, holding dataset, split, augmentation, loss, optimiser, and schedule fixed while varying only the encoder and the presence of attention gates, so that the reported deltas are attributable. Second, we propose a Multi-Factor Severity Index (MFSI) that folds lesion count, boundary irregularity, and spatial dispersion into the severity score alongside area, and we validate it against ground truth. Third, we characterise robustness under three corruption types at three intensities each. Fourth, we identify a data leakage pattern in prior work on this dataset and quantify how much of the apparent performance gap it explains.

Section II reviews related work. Section III describes the dataset, architectures, loss, and severity index. Section IV reports results. Section V discusses what the numbers mean, including the negative result on attention gates. Section VI concludes.

II. RELATED WORK

A. From handcrafted features to learned ones

Early plant disease systems extracted colour histograms in HSV or CIELAB space, texture descriptors from gray-level co-occurrence matrices, and shape moments, then fed these to an SVM or random forest [7]. Singh and Misra [8] combined image segmentation with a genetic algorithm on 106 images and reported 95.6% accuracy. Elangovan and Nalini [9] paired Otsu thresholding and k-means clustering with an SVM. These pipelines worked in the conditions they were tuned for and broke elsewhere, because the features encoded assumptions about lighting and leaf morphology that the designer had baked in by hand.

B. Deep classification

AlexNet [10] made learned features the default. VGG [11], ResNet [12], and EfficientNet [13] followed, and each was applied to PlantVillage [14] in turn. Ferentinos [4] benchmarked several backbones and reported accuracies above 99% on held-out PlantVillage images, though the dataset's uniform laboratory backgrounds make those figures difficult to transfer to field conditions. Madhurya and Jubilson [15] proposed YR2S, a lightweight classifier reaching 98.1% on a 12,000-image benchmark. All of this work outputs a label.

C. Segmentation architectures in agriculture

Long et al. [16] introduced fully convolutional networks for dense prediction, and Ronneberger et al. [5] refined the idea into U-Net. Agricultural adaptations followed quickly. Yuan et al. [17] built a spatial pyramid encoder-decoder cascade for crop disease leaves. Wang et al. [18] proposed MFBP-UNet for pear leaf disease under natural field backgrounds. Deng et al. [19] developed MC-UNet for tomato leaves. Zhang and Zhang [20] modified U-Net for diseased leaf segmentation on a PlantVillage subset. El-Assiouti et al. [21] paired a lightweight U-Net with a super-resolution GAN to segment small lesions in low-resolution inputs. Abinaya et al. [22] combined an attention residual U-Net with a cascading autoencoder for multi-class segmentation.

Attention mechanisms recur throughout this line of work, and the reported gains are usually positive but small. This is worth holding onto when interpreting our own attention result in Section V.

D. Severity estimation

Bock et al. [2] documented the variability of manual severity assessment and argued for image-based quantification. Esgario et al. [23] trained a multi-task network to classify coffee leaf biotic stress and estimate its severity jointly, which is the closest prior work in spirit to the MFSI, though their severity head is learned rather than computed from a segmentation mask. Most segmentation papers that mention severity at all compute the pixel ratio and move on.

E. Interpretability and robustness

Selvaraju et al. [24] introduced Grad-CAM, which produces a heatmap of the image regions whose activations most influence a prediction. Gal and Ghahramani [25] framed dropout at inference as approximate Bayesian inference, giving a cheap route to predictive uncertainty. Hendrycks and Dietterich [26] proposed a standard corruption benchmark for image classifiers. None of these three tools is standard practice in plant disease segmentation papers, which is a gap this work partially closes.

F. The gap

Existing work either classifies without localising, localises without quantifying severity beyond a pixel ratio, or reports accuracy without checking whether the model degrades gracefully or focuses on the right pixels. We address all four in one pipeline, on a dataset small enough that the compute fits inside a free-tier Colab session.

3.3 Theoretical framework

Variables of the study contains dependent and independent variable. The study used pre-specified method for the selection ofvariables. The study used the Stock returns are as dependent variable. From the share price of the firm the Stock returns are calculated. Rate of a stock salable at stock market is known as stock price.

III. METHODOLOGY

A. Dataset

We use the Leaf Disease Segmentation dataset published on Kaggle [27], which contains 588 RGB photographs of plant leaves with pixel-level binary masks marking diseased tissue. The images span several crops and symptom types including bacterial spot, early blight, and leaf mould, and the masks carry no disease-class label: the task is diseased against healthy, not which disease. Mask pixels encoding disease are stored in red, with BGR value (128, 0, 0).

That encoding detail matters more than it looks. Decoding a mask to single-channel grayscale applies the luminance weighting 0.299R + 0.587G + 0.114B, which maps the disease value 128 to 38 out of 255, or 0.149 after normalisation. Any threshold at 0.5 then erases every annotated lesion, and the model trains on empty masks while reporting high pixel accuracy for predicting healthy everywhere. We decode masks as three-channel RGB, take the per-pixel maximum across channels, and threshold at 30 out of 255 to tolerate JPEG artefacts.

Fig. 1. Representative dataset samples. Left: input leaf image. Centre: ground truth mask, with the diseased fraction of the frame annotated. Right: mask overlaid on the input in red.

Diseased area across the dataset ranges from 0.4% to 63.8% of the frame, with a mean of 16.9% and a median of 12.9%. The distribution is right-skewed: most leaves carry modest infection, a minority are heavily affected. Fig. 2 shows the histogram.

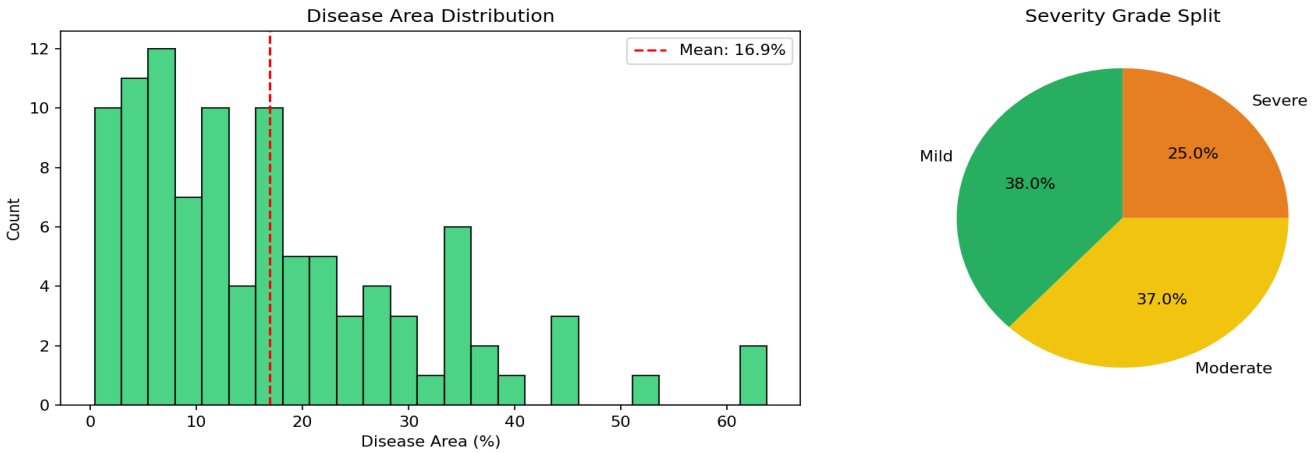

Fig. 2. Distribution of diseased area across the dataset, with the severity grade split derived from the area thresholds in Table I.

B. Split protocol and the leakage question

The Kaggle release ships a pre-augmented folder containing 2,940 images, five variants per original. Training on that folder and then splitting is the obvious shortcut and it is a mistake, because a horizontal flip of leaf 042 in the validation fold is not an unseen sample when the original leaf 042 sat in the training fold. The model has already memorised the lesion. Reported validation IoU rises accordingly and means nothing about field performance.

We therefore split the 588 originals first, with a fixed seed, into 470 training and 118 validation pairs, and apply augmentation only to the training fold, generated fresh each epoch. Section V returns to this point because it explains most of the gap between our numbers and the published state of the art on this dataset.

C. Augmentation

Augmentation runs on the fly through the Albumentations library, which guarantees that a geometric transform applied to an image is applied identically to its mask. Twelve transforms are configured. Geometric: horizontal flip (p = 0.5), vertical flip (p = 0.3), 90-degree rotation (p = 0.5), shift-scale-rotate with limits of 10% translation, 20% scale, and 25 degrees (p = 0.5), elastic deformation with alpha 80 and sigma 40 (p = 0.3), and random resized crop retaining 75% to 100% of the frame (p = 0.3). Photometric, applied to the image only and selected one of three: brightness-contrast jitter at 25%, CLAHE with clip limit 3.0, or hue-saturation-value shift (p = 0.7 for the group). Degradation: Gaussian noise (p = 0.2), Gaussian blur with kernel 3 to 5 (p = 0.2), synthetic shadow (p = 0.2). Regularisation: coarse dropout of one to four rectangular holes (p = 0.2).

Elastic deformation deserves a note. Leaves bend and curl; a lesion on a flat leaf and the same lesion on a curled one are the same lesion. Elastic warping teaches that invariance in a way that rigid rotation cannot.

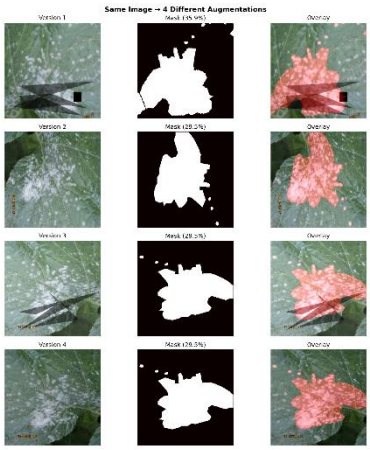

Fig. 3. Four independently sampled augmentations of a single training image. The mask tracks every geometric transform applied to the image.

D. Architectures

Three models are trained. The baseline is a vanilla U-Net with a four-level encoder (32, 64, 128, 256 filters), a 512-filter bottleneck, and a mirrored decoder, initialised from scratch. The second replaces the encoder with EfficientNet-B0 [13] pretrained on ImageNet, tapping four intermediate activations as skip connections at 128, 64, 32, and 16 pixel resolution plus the 8-pixel top activation as the bridge. The third is identical to the second except that each skip connection passes through an attention gate before concatenation.

The attention gate takes the encoder skip tensor and the decoder gating tensor from the level below. The gating tensor is resized to the skip resolution, both are projected through 1x1 convolutions into a shared channel space, summed, passed through ReLU, projected to a single channel, and squashed by a sigmoid to produce a spatial weight map. The skip tensor is multiplied elementwise by that map. Regions the decoder considers irrelevant are attenuated toward zero; regions it considers promising pass through intact.

[INSERT FIGURE 4 HERE] file: SUGGESTED FIGURE — architecture diagram (not yet generated)

Fig. 4. Proposed architecture. EfficientNet-B0 encoder, attention-gated skip connections, U-Net decoder, sigmoid output, and the MFSI post-processing stage. Suggested Figure.

E. Loss

Healthy pixels outnumber diseased pixels by roughly five to one on average in this dataset, and the imbalance is worse on lightly infected leaves. Plain binary cross-entropy rewards a model that predicts healthy everywhere. We combine Dice loss, which measures relative overlap and is therefore insensitive to the size of the positive class, with focal loss, which downweights easy pixels and concentrates gradient on the ambiguous ones at lesion boundaries.

$$L_{dice} = 1 - (2 * sum(y * p) + eps) / (sum(y) + sum(p) + eps) \quad (1)$$

$$L_{focal} = -alpha * (1 - p_t)^{gamma} * log(p_t) \quad (2)$$

$$L = 0.6 * L_{dice} + 0.4 * L_{focal} \quad (3)$$

with alpha = 0.25, gamma = 2.0, and eps = 1e-6. The 0.6/0.4 weighting keeps overlap as the primary objective while leaving enough focal signal to sharpen boundaries.

F. Multi-Factor Severity Index

Given a predicted binary mask M, the MFSI computes four normalised sub-scores.

Area. The diseased fraction, normalised against a 50% ceiling beyond which the leaf is effectively lost:

$$S_{area} = \min((\text{sum}(M) / |M|) / 0.50 , 1) \quad (4)$$

Lesion count. Connected component labelling yields n distinct lesions. Multiple separate lesions indicate either repeated inoculation or systemic spread, both worse prognoses than a single focus of equal area:

$$S_{count} = \min(n / 5 , 1) \quad (5)$$

Boundary complexity. For the largest lesion contour with area A and perimeter P, circularity is 4*pi*A / P^2, equal to 1 for a disc and falling as the outline becomes ragged. Ragged margins are associated with actively expanding necrosis:

$$S_{shape} = 1 - \min(4*\pi*A / P^2 , 1) \quad (6)$$

Spatial dispersion. Let (y_i, x_i) be the centroids of the n lesions. The normalised standard deviation of centroid position measures whether disease is clustered or scattered across the blade:

$$S_{spread} = \min(\sqrt{ sd(x)^2 + sd(y)^2 } / 0.35 , 1) \quad (7)$$

The index is their weighted sum, scaled to 0 to 100:

$$MFSI = 100 * (0.4*S_{area} + 0.2*S_{count} + 0.2*S_{shape} + 0.2*S_{spread}) \quad (8)$$

Area keeps the dominant weight because absolute tissue loss remains the primary driver of yield impact. The other three carry equal minority weights and can shift a borderline case one grade in either direction. Weights and normalisation constants were set by inspection of the dataset and were not tuned against the validation fold; making them learnable is left to future work, and we say plainly that this is a limitation.

TABLE I. SEVERITY GRADE THRESHOLDS

Grade	MFSI range	Interpretation
Healthy	0 - 5	No actionable infection
Mild	5 - 25	Monitor; localised early lesions
Moderate	25 - 50	Treat; remove worst-affected tissue
Severe	50 - 100	Immediate intervention; consider isolation

G. Training setup

All three models are trained with Adam, batch size 8, on 256 by 256 inputs, using cosine annealing with a five-epoch linear warmup and a floor of 1e-6. The two pretrained models follow a two-stage protocol: ten epochs with the encoder frozen at a peak learning rate of 1e-3, so that the randomly initialised decoder does not push destructive gradients into ImageNet weights, then fine-tuning of the full network at 1e-4. Early stopping monitors validation IoU with patience 15. Training ran on a single NVIDIA T4 in Google Colab under TensorFlow 2.19 with Keras 3. Total wall-clock time for all three models was approximately one hour.

H. Inference

At test time we average predictions over four flips of the input (identity, horizontal, vertical, both) and reverse each flip on the corresponding output before averaging. Predictions are thresholded at 0.5, then cleaned by morphological closing with a 5x5 elliptical kernel, opening with the same kernel, and removal of connected components smaller than 50 pixels.

IV. RESULTS

A. Segmentation

Table II reports the three models on the 118-image validation fold. Every metric is computed with test-time augmentation and morphological post-processing applied uniformly.

TABLE II. SEGMENTATION PERFORMANCE ON THE VALIDATION FOLD (N = 118)

Model	Params	IoU	Dice	Acc.	Prec.	Rec.	F1
Vanilla U-Net	7.85 M	0.586	0.709	0.911	0.706	0.784	0.743
EfficientNet U-Net	9.01 M	0.681	0.788	0.943	0.822	0.834	0.828
Attn U-Net (ours)	9.40 M	0.670	0.779	0.944	0.871	0.779	0.822

The pretrained encoder is responsible for almost all of the improvement over the from-scratch baseline: 0.095 IoU, a relative gain of 16.1%. Adding attention gates does not improve IoU. It lowers it by 0.010 while raising precision by 0.049 and dropping recall by 0.055. The attention model is more conservative about calling a pixel diseased, and on this dataset that trade is roughly

neutral in F1 terms. We treat this as a negative result and discuss it in Section V rather than presenting the small accuracy uptick (0.9429 to 0.9445) as if it settled the question.

TABLE III. ABLATION: INCREMENTAL CONTRIBUTION OF EACH COMPONENT

Configuration	IoU	delta IoU	Precision	Recall
U-Net, no TL, no attention	0.586	—	0.706	0.784
+ ImageNet transfer learning	0.681	+0.095	0.822	0.834
+ attention gates	0.670	-0.010	0.871	0.779

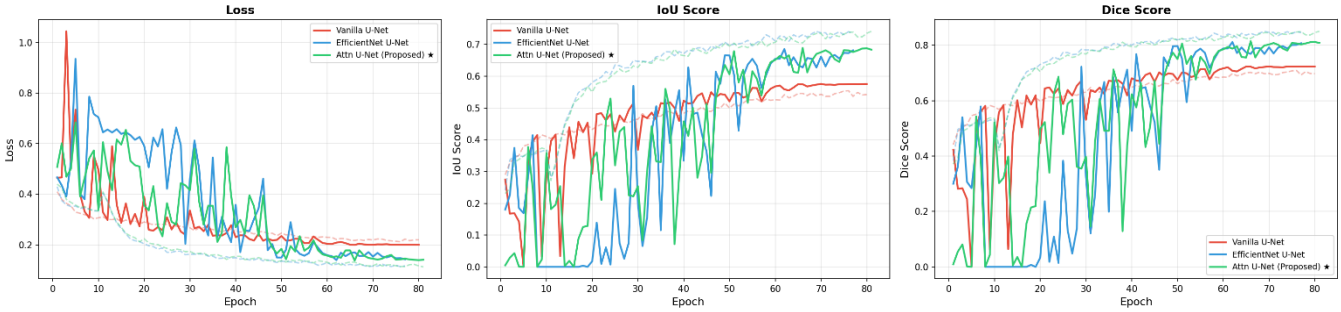

Fig. 5. Training and validation curves for loss, IoU, and Dice across the three models. Dashed lines are training, solid lines validation. Both pretrained models converge to a lower loss floor than the from-scratch baseline, and the attention variant tracks the non-attention variant closely throughout.

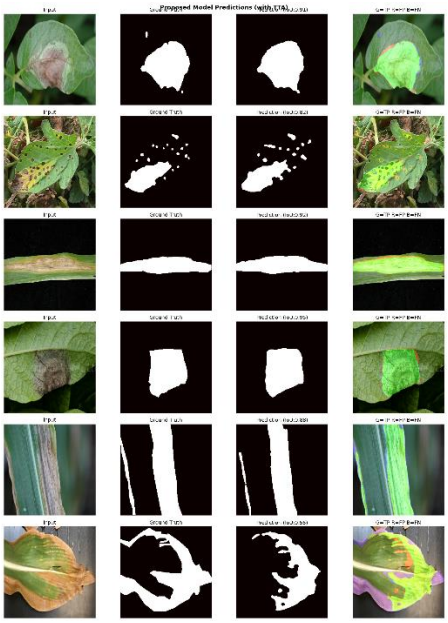

Fig. 6. Qualitative predictions from the attention model with TTA. Columns: input, ground truth, prediction with per-image IoU, and an error overlay where green is true positive, red is false positive, and blue is false negative. Per-image IoU across these samples ranges from 0.55 to 0.95.

B. Severity estimation

MFSI-derived diseased area correlates with the ground truth area at Pearson r = 0.875 and Spearman rho = 0.891 across the 118 validation images, with a mean absolute error of 3.86 percentage points. Grade assignment is correct for 95 of 118 images, or 80.5%.

TABLE IV. SEVERITY GRADE CONFUSION MATRIX (ROWS = TRUE, COLUMNS = PREDICTED)

True \ Pred	Healthy	Mild	Moderate	Severe	Total
Healthy	0	0	0	0	0
Mild	1	41	5	0	47
Moderate	0	8	38	2	48
Severe	0	1	6	16	23

The error structure is informative. Of the 23 misgraded images, 21 fall into an adjacent grade and only two skip a grade. Nothing labelled Severe was called Mild, and nothing labelled Mild was called Severe. The single largest error cluster is eight Moderate leaves graded Mild, which traces to under-segmentation of faint lesion margins: when the model misses the diffuse edge of a lesion,

both the area term and the boundary complexity term fall. No validation image carries a Healthy ground truth grade, since every image in this dataset contains disease, so the Healthy row is empty by construction and the grade classifier is effectively a three-way problem here.

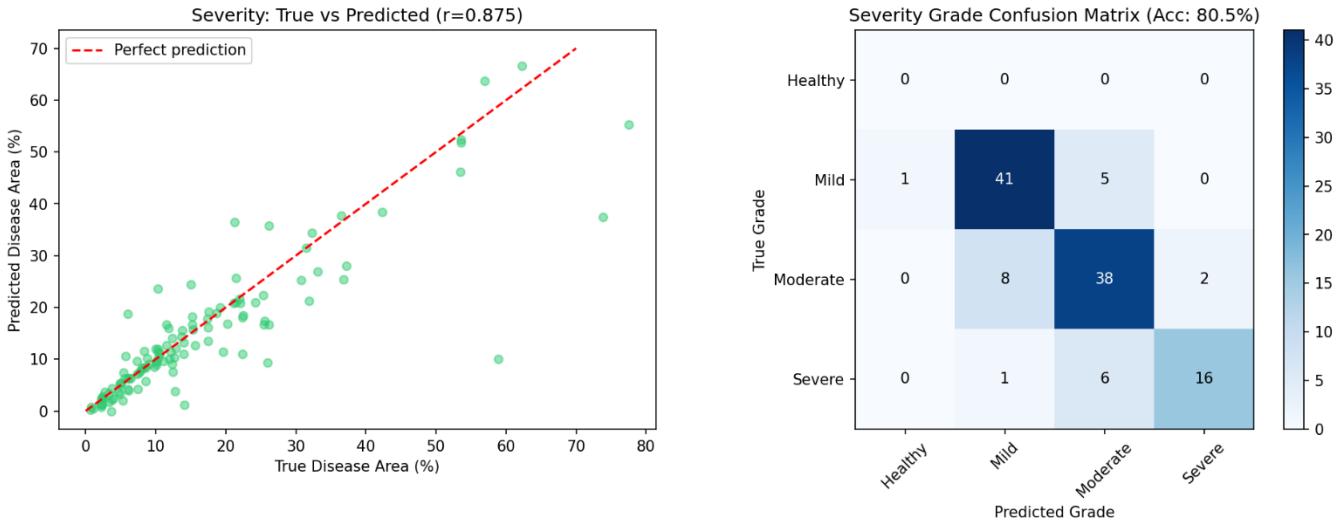

Fig. 7. Left: predicted against ground truth diseased area, with the identity line. Right: severity grade confusion matrix. Points cluster along the diagonal, with dispersion widening above 40% diseased area where few training examples exist.

C. Robustness

Table V reports IoU under three corruption families at three intensities, evaluated on 50 validation images.

TABLE V. IOU UNDER SYNTHETIC CORRUPTION

Condition	IoU	Change vs clean
Clean	0.698	—
Gaussian noise, level 1	0.695	-0.003
Gaussian noise, level 2	0.665	-0.033
Gaussian noise, level 3	0.557	-0.141
Blur, level 1	0.666	-0.032
Blur, level 2	0.620	-0.079
Blur, level 3	0.532	-0.166
Brightness shift, level 1	0.680	-0.019
Brightness shift, level 2	0.615	-0.083
Brightness shift, level 3	0.452	-0.246

Level-1 corruption of any kind costs at most 0.032 IoU, which is within the range a field deployment could absorb. Level 3 is a different matter, and brightness is the worst offender: a strong exposure shift removes 0.246 IoU, more than a third of clean performance. This is the corruption our augmentation pipeline covers least aggressively (brightness jitter was capped at 25%, while the level-3 test applies a 45% shift), and the result is a fair criticism of that choice rather than an inherent property of the architecture.

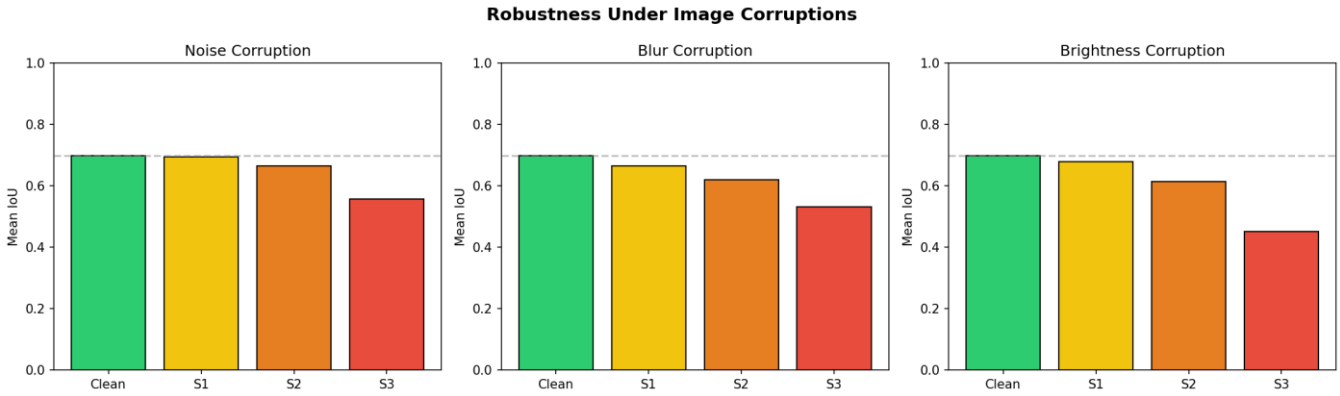

Fig. 8. Mean IoU under noise, blur, and brightness corruption at three intensities. The dashed line marks clean performance.

D. Interpretability

Grad-CAM maps computed on the attention model concentrate on lesion tissue rather than on leaf veins, background soil, or the human hands visible in several dataset images. This does not prove the model reasons correctly, but it rules out the failure mode where a segmentation network keys on a spurious background correlate.

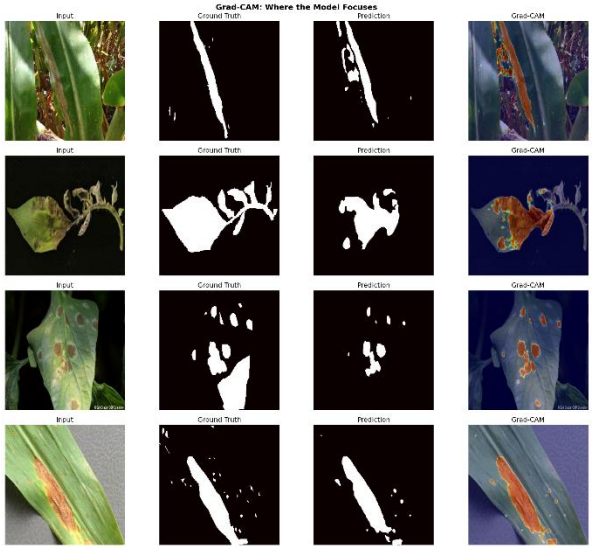

Fig. 9. Grad-CAM activation maps for four validation images. Columns: input, ground truth, prediction, and Grad-CAM overlay. Activation concentrates on lesion tissue in all four cases.

V. DISCUSSION

A. Why attention did not help, and what it did instead

The attention gates add 391,524 parameters and lose 0.010 IoU. Two explanations fit the evidence. The first is sample size: 470 training images is thin for learning a spatial weighting function on top of an encoder that already produces good features, and the gate has no strong gradient signal telling it which regions to suppress when the decoder is already competent. The second is that the gate is doing something real but different from what IoU measures. Precision rises from 0.822 to 0.871, a 6% relative improvement, while recall falls by a comparable margin. The gate is suppressing skip-connection features in ambiguous regions, which removes false positive lesion pixels and simultaneously removes some true ones.

Which behaviour is preferable depends on the deployment. A spot-spraying robot that acts on every predicted lesion pixel pays for false positives in wasted chemical and should prefer the attention model. A screening tool meant to flag leaves for human review pays for false negatives in missed disease and should prefer the non-attention model. We report both rather than declaring a winner, and we note that a paper reporting only accuracy (0.9429 against 0.9445) would have concluded that attention helped.

B. The gap against published results on this dataset

Singh et al. [28] use the same 588-image Kaggle dataset and report 0.943 validation IoU with 98.99% accuracy from a U-Net with attention mechanisms and identity mappings. Our best IoU is 0.681. That is not a small gap and it deserves an explanation rather than a hedge.

Their methodology, as described in their Section 3.1 and 3.5.3, augments the 588 originals to 7,056 images and then applies a 70:30 train-validation split to the augmented set. Under that procedure, a flipped or brightness-shifted copy of a given leaf can land in validation while the original sits in training. The network has already fit that lesion; scoring it again as validation measures memorisation, not generalisation. Their reported numbers are internally consistent and their architecture is sound, but the split makes the validation IoU an unreliable estimate of field performance. The same criticism applies to any of the comparison figures in their Table 4 that follow the same protocol, which we cannot verify from primary sources and therefore do not tabulate here.

We do not claim our 0.681 is the ceiling for this dataset. We claim it is an honest estimate under a leak-free split, and that any future work on this benchmark should state its split protocol explicitly enough that a reader can tell which regime it is in.

TABLE VI. COMPARISON WITH PRIOR WORK ON THE SAME DATASET

Work	Split protocol	Images trained on	IoU	Accuracy
Singh et al. [28]	Augment to 7056, then split 70:30	4,939	0.943	0.9899
This work (Attn U-Net)	Split 588 originals, then augment train fold only	470	0.670	0.9445
This work (EffNet U-Net)	Split 588 originals, then augment train fold only	470	0.681	0.9429

The comparison in Table VI is not apples to apples, and that is the point of including it.

C. Transfer learning carries the result

The 0.095 IoU gain from swapping a scratch encoder for an ImageNet-pretrained EfficientNet-B0 is the largest single effect in the study, and it costs only 1.15 M additional parameters. On a dataset of this size the encoder simply cannot learn good low-level features from 470 images; it borrows them. Any practitioner with a small agricultural dataset should exhaust transfer learning before reaching for architectural novelty, which is a boring conclusion but appears to be the correct one.

D. Severity index

An r of 0.875 and 80.5% grade agreement compare favourably against the inter-rater reliability of human visual assessment, which Bock et al. [2] place below kappa 0.6 in many trials. The index's failure mode is systematic rather than random: it under-grades leaves whose lesions have soft margins, because under-segmentation shrinks both the area term and the shape term at once. That coupling is a design weakness. Decoupling them, or estimating boundary complexity from the model's soft probability map rather than the thresholded mask, is a concrete next step.

E. Limitations

Each model was trained once. We report no variance across seeds and cannot make claims about the significance of the 0.010 IoU difference between the two pretrained models, which may well sit inside run-to-run noise. The dataset is binary and small; the system says diseased, not which disease, which limits its clinical usefulness. MFSI weights and thresholds are hand-set. Robustness was tested against synthetic corruptions, not real field imagery. And the validation fold contains no healthy leaves, so we have no evidence about false positive behaviour on uninfected tissue, which is arguably the most important failure mode for a spraying application.

VI. CONCLUSION AND FUTURE WORK

We trained three segmentation models on the Kaggle Leaf Disease Segmentation dataset under a leak-free split and a single controlled protocol. Transfer learning from ImageNet delivered a 0.095 IoU improvement over a from-scratch U-Net. Attention gates did not improve IoU; they traded 0.055 recall for 0.049 precision, which is useful in some deployments and harmful in others. A Multi-Factor Severity Index combining diseased area, lesion count, boundary circularity, and spatial dispersion reached Pearson r = 0.875 against ground truth severity and graded 80.5% of validation leaves correctly, with errors confined almost entirely to adjacent grades. The model degrades gently under mild corruption and sharply under severe brightness shift, a weakness traceable to our augmentation configuration.

We also documented a leakage pattern in the leading published result on this dataset, and we would encourage the field to report split protocols with enough precision that readers can tell whether a validation number means anything.

Four directions follow. Repeat every experiment across multiple seeds and report variance, without which the attention comparison remains inconclusive. Retrain on PlantSeg [29], which offers over 11,000 segmented images across 115 disease types, and test whether attention gates earn their parameters given more data. Learn the MFSI weights against expert-annotated severity scales instead of setting them by hand. Add uncertainty estimation via Monte Carlo dropout so that low-confidence predictions can be routed to a human rather than acted on.

REFERENCES

[1] S. Savary, L. Willocquet, S. J. Pethybridge, P. Esker, N. McRoberts, and A. Nelson, "The global burden of pathogens and pests on major food crops," Nature Ecology & Evolution, vol. 3, no. 3, pp. 430-439, Mar. 2019, doi: 10.1038/s41559-018-0793-y.

[2] C. H. Bock, P. E. Parker, A. Z. Cook, and T. R. Gottwald, "Visual rating and the use of image analysis for assessing different symptoms of citrus canker on grapefruit leaves," Plant Disease, vol. 94, no. 7, pp. 845-854, Jul. 2010, doi: 10.1094/PDIS-94-7-0845.

[3] S. P. Mohanty, D. P. Hughes, and M. Salathe, "Using deep learning for image-based plant disease detection," Frontiers in Plant Science, vol. 7, p. 1419, Sep. 2016, doi: 10.3389/fpls.2016.01419.

[4] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Computers and Electronics in Agriculture, vol. 145, pp. 311-318, Feb. 2018, doi: 10.1016/j.compag.2018.01.009.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 2015, pp. 234-241, doi: 10.1007/978-3-319-24574-4_28.

[6] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, Apr. 2018.

[7] S. Arivazhagan, R. N. Shebiah, S. Ananthi, and S. V. Varthini, "Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features," Agricultural Engineering International: CIGR Journal, vol. 15, no. 1, pp. 211-217, Mar. 2013.

[8] V. Singh and A. K. Misra, "Detection of plant leaf diseases using image segmentation and soft computing techniques," Information Processing in Agriculture, vol. 4, no. 1, pp. 41-49, Mar. 2017, doi: 10.1016/j.inpa.2016.10.005.

[9] K. Elangovan and S. Nalini, "Plant disease classification using image segmentation and SVM techniques," International Journal of Computational Intelligence Research, vol. 13, no. 7, pp. 1821-1828, 2017.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems (NeurIPS), Lake Tahoe, NV, USA, 2012, vol. 25, pp. 1097-1105.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. on Learning Representations (ICLR), San Diego, CA, USA, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[13] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. on Machine Learning (ICML), Long Beach, CA, USA, 2019, pp. 6105-6114.

[14] D. P. Hughes and M. Salathe, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," arXiv preprint arXiv:1511.08060, Nov. 2015.

[15] C. Madhurya and E. A. Jubilson, "YR2S: Efficient deep learning technique for detecting and classifying plant leaf diseases," IEEE Access, vol. 12, pp. 3790-3804, 2024, doi: 10.1109/ACCESS.2023.3345583.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 3431-3440, doi: 10.1109/CVPR.2015.7298965.

[17] Y. Yuan, Z. Xu, and G. Lu, "SPEDCCNN: Spatial pyramid-oriented encoder-decoder cascade convolution neural network for crop disease leaf segmentation," IEEE Access, vol. 9, pp. 14849-14866, 2021, doi: 10.1109/ACCESS.2021.3052769.

[18] H. Wang, J. Ding, S. He, et al., "MFBP-UNet: A network for pear leaf disease segmentation in natural agricultural environments," Plants, vol. 12, no. 18, p. 3209, Sep. 2023, doi: 10.3390/plants12183209.

[19] Y. Deng, H. Xi, G. Zhou, et al., "An effective image-based tomato leaf disease segmentation method using MC-UNet," Plant Phenomics, vol. 5, p. 49, 2023, doi: 10.34133/plantphenomics.0049.

[20] S. Zhang and C. Zhang, "Modified U-Net for plant diseased leaf image segmentation," Computers and Electronics in Agriculture, vol. 204, p. 107511, Jan. 2023, doi: 10.1016/j.compag.2022.107511.

[21] H. S. El-Assiouti, H. El-Saadawy, M. N. Al-Berry, and M. F. Tolba, "Lite-SRGAN and Lite-UNet: Toward fast and accurate image super-resolution, segmentation, and localization for plant leaf diseases," IEEE Access, vol. 11, pp. 67498-67517, 2023, doi: 10.1109/ACCESS.2023.3289750.

[22] S. Abinaya, K. U. Kumar, and A. S. Alphonse, "Cascading autoencoder with attention residual U-Net for multi-class plant leaf disease segmentation and classification," IEEE Access, vol. 11, pp. 98153-98170, 2023, doi: 10.1109/ACCESS.2023.3312718.

[23] J. G. M. Esgario, R. A. Krohling, and J. A. Ventura, "Deep learning for classification and severity estimation of coffee leaf biotic stress," Computers and Electronics in Agriculture, vol. 169, p. 105162, Feb. 2020, doi: 10.1016/j.compag.2019.105162.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

[25] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. Int. Conf. on Machine Learning (ICML), New York, NY, USA, 2016, pp. 1050-1059.

[26] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in Proc. Int. Conf. on Learning Representations (ICLR), New Orleans, LA, USA, 2019.

[27] Fakhrealam9537. "Leaf disease segmentation dataset." Kaggle. https://www.kaggle.com/datasets/fakhrealam9537/leaf-disease-segmentation-dataset (accessed Jul. 13, 2026).

[28] G. Singh, A. A. Al-Huqail, A. Almogren, et al., "Enhanced leaf disease segmentation using U-Net architecture for precision agriculture: A deep learning approach," Food Science & Nutrition, vol. 13, no. 7, p. e70594, Jul. 2025, doi: 10.1002/fsn3.70594.

[29] T. Wei, Z. Chen, X. Yu, S. Chapman, A. Melloy, and Z. Huang, "PlantSeg: A large-scale in-the-wild dataset for plant disease segmentation," arXiv preprint arXiv:2409.04038, Sep. 2024.