



Development of an RF-DETR Based Model for Sugarcane Leaf Disease Detection Using Deep Learning Techniques

Madhulika Mittal*¹ and Md. Iqbal²

^{1,2} Quantum University, Roorkee, India

*¹madhulikamittal24@gmail.com, ²iqbal.hodcse@gmail.com

Abstract. Red rot, rust, leaf scald and yellow leaf syndrome are the four major foliar diseases of sugarcane (*Saccharum Officinarum*), which is a major cash crop in a critical situation throughout the world. Most of the conventional detection methods rely on visual inspection by agronomists, which is labour-intensive and inconsistent. By combining a convolutional backbone for capturing local spatial features with a transformer-based encoder-decoder for long-range contextual modelling, this study proposes RF-DETR, a novel hybrid architecture that achieves superior performance. The key feature of RF-DETR is its learnable region-focused attention module that dynamically filters out irrelevant background regions and enhances discriminative disease-related patches. The focused attention allows accurate bounding-box localisation of lesions even in challenging field conditions. The model is trained using 4200 annotated images across five different categories (red rot, rust, leaf scald, yellow leaf disease, healthy leaves) from agricultural fields and public repositories. A standard object-detection metrics reports mAP@0.5 of 93.1%, mAP@0.5:0.95 of 81.4%, and a mean IoU of 87.6% evaluated using experiments. The suitability of the RF-DETR in precision agriculture applications makes it more competitive than the baseline architectures such as CNN, ResNet-50, and Vision Transformer.

Keywords: Sugarcane Disease Detection · RF-DETR · Region-Focused Attention · Deep Learning · Precision Agriculture · Object Detection · Transformer

1. Introduction

Sugarcane is grown in over 100 countries in the tropical or subtropical regions, and produces about 80% of the world's sugar. In addition to sugar, the crop also provides raw material for bioethanol, paper and bioplastics and plays a crucial role in food security and rural economy. When untreated, however, foliar diseases consistently affect harvestable crop losses of 10–30%, and in vulnerable areas can result in almost complete crop losses. Conventional disease monitoring is based on occasional field scouting by trained agronomists. It is time consuming, suffers from inter-observer variability, and is impractical for large continuous monocultures as found in commercial sugarcane production. Disease is often not detected until pathogen levels are high enough to cause economic losses. With the advent of deep learning, plant pathology diagnostics has changed significantly. Convolutional neural networks (CNNs) operate with hierarchical spatial filters, to identify normal tissue from abnormal lesions with high fidelity. However, conventional CNN processes images via pre-defined receptive fields and fails to learn the

long-range spatial relationship of diseases that reflects the spread patterns of disease throughout the whole leaf surface.

This limitation is addressed by transformer architectures, which model dependencies without considering spatial distance. However, generic transformers spread their attention equally over all image tokens, thus distributing attention over the small and irregularly shaped lesion regions typical of early-stage fungal and bacterial infections.

To address this imbalance, RF-DETR proposes a learnable region-focused attention module between the convolutional backbone and the transformer encoder. This module generates a spatial saliency map that boosts the tokens representing candidate disease patches while reducing the homogeneous background tokens before they are fed into the self-attention layers. This yields a detection model which is both lesion texture-sensitive (as captured in CNN features) and contextually aware (as captured in transformer attention) of the spatial distribution of lesions on the leaf. Contributions of this work can be listed below:

- Novel architecture RF-DETR by integrating ResNet-50 features with a region-based attention model and an encoder-decoder transformer for detection heads.
- The collection of publicly available 4,200 sugarcane leaves annotated with bounding box details into five diseases.
- Extensive evaluation based on $mAP@0.5$, $mAP@0.5:0.95$, mean IoU, and classification metrics [1].
- Ablation study to quantify the role of the region-based attention model [2].
- Performance comparison with CNN, ResNet-50, ViT, YOLOv11, and EfficientDet models.

2. Related Work

Initial attempts at computational methods for plant disease detection used manually engineered feature descriptions and were applied using classical Random Forests and SVM classifiers. While yielding adequate results, these techniques were affected negatively by variations in lighting conditions in the case of real-life datasets. Presence of background noise, apparatus noise and disease form [3]. In order to take the development in this area a step further, CNN architectures which enabled learning hierarchies of feature representations from raw pixels were utilized, without the necessity for hand-designed feature extraction [4]. Architecture such as VGG, Inception, ResNet and Densenet. VGG, Inception, ResNet and Densenet proved that, through sufficient depth and residual connections, it is possible to reach an accuracy comparable to human perception. In the large-scale classification tests performed on ImageNet dataset, while being easily transferrable to agro-biological domain-specific cases. Research conducted on the diseases of rice blast, tomato leaf mold and maize northern blight found that CNNs were superior to agro-economists in their ability of classification accuracy.

The development of the vision transformer-based models brought about paradigm shift in recent years through replacing the spatial convolution operations with multi-head attention.self-attention. The idea of Vision Transformer (ViT) was that a sequence of tokens can be extracted from the patches of images, and when enough training data is provided, the stack of attention layers can do as good as or even better than the CNN [5]. Anchors have been eliminated by using detection-oriented transformers such as DETR (Detection Transformer). Boxes and non-maximum suppression (NM) – Object detection as a problem of predicting the set of objects via bipartite graph.matching. To improve the convergence process, deformable DETR allows for only sparse, learned reference points [1]. In the agriculture field, the use of YOLO variations has been found to be useful in detecting plants affected with diseases due to its capability to predict the appearance of the disease in real time. YOLOv11 variations have found wide usage in agriculture for real-time plant disease detection since they can detect the appearance of the disease [1]. Good speed accuracy tradeoff. YOLOv11 actually produces competitive mAP on publicly available datasets.Process large amounts of plant disease data at inference rates of drone cameras. EfficientDet This work builds upon the concept of efficiency by proposing compound scaling of multi-scale feature pyramids [6]. However, an important missing component None of these designs account for a strategy for minimizing background noise. and drawing attention to lesion regions that are smaller and have irregular shapes in sugarcane imagery with less than 5% lesion coverage.

3. Suggested Method

3.1 System Outline

The overall pipeline of RF-DETR consists of four stages: (1) multi-scale features are extracted using a ResNet-50 backbone with a feature pyramid, (2) Regions' spatial saliency is estimated using a region-focused attention module, (3) Global context is modelled with a transformer encoder-decoder, and (4) Global class prediction and bounding-box regression are performed by a detection head. The output of the system are triplets of the form (class label, bounding box, confidence score) corresponding to each disease lesion detected.

3.2 Dataset

A dedicated dataset was assembled to support training and evaluation. Images were sourced from two streams: (i) direct field collection at commercial sugarcane farms in Uttar Pradesh, India, during the 2023 growing season, and (ii) validated public repositories including PlantVillage and the Sugarcane Disease Dataset hosted on Kaggle. Field images were captured under natural illumination using smartphones (12 MP rear cameras) and a consumer-grade DSLR at image resolutions ranging from 2 MP to 16 MP before standardisation.

The compiled dataset comprises 4,200 images distributed across five categories as follows:

Table 1. Dataset Distribution Across Disease Categories and Image Sources

Disease Category	Field Images	Public Images	Total	% of Dataset
Red Rot	430	420	850	20.2%
Rust	390	460	850	20.2%
Leaf Scald	350	490	840	20.0%
Yellow Leaf	370	450	820	19.5%
Healthy	410	430	840	20.0%
Total	1,950	2,250	4,200	100%

BBox annotation has been used for annotations in LabelImg following Pascal VOC XML format. Two domain experts have independently annotated each image, and Cohen's Kappa ($\kappa = 0.89$) has revealed near-perfect agreement between the annotators. Conflicts between the annotations have been resolved by a senior plant pathologist [7]. To guarantee balanced classes in the data set, stratified random sampling has been performed so that 70%, 15%, and 15% of the images are included in training, validation, and test sets, respectively. Each image in the field data set does not appear more than once in the partitions. To examine the consistency of the models, five training sessions have been executed using various seeds, and the average and standard deviation across all runs are presented [8].

3.3 Preprocessing and Data Augmentation

- Any images obtained were rescaled to 512×512 pixels [9]. Channel-wise normalization was done by taking the mean values from ImageNet as $\mu = [0.485, 0.456, 0.406]$ and standard deviation values $\sigma = [0.229, 0.224, 0.225]$, following backbone parameters. Selective Gaussian blur ($\sigma=1.0$) was adopted for filtering out any sensor noise while keeping intact lesion boundaries. Data augmentation was carried out only on the training set to avoid any data leakage. Data augmentation techniques include:
- Random horizontal and vertical flipping ($p = 0.5$)
- Random rotation within $[-15^\circ, +15^\circ]$ using edge padding [10]
- Color jitter: brightness (+-)0.3, contrast (+-)0.3, saturation (+-)0.2, hue (+-)0.05 [11]
- Random Zoom with scale factor in range $[0.8, 1.2]$

- Mosaic augmentation: four training images tiled into a single composite to expose the model to multi-scale disease context

Bounding-box coordinates were transformed consistently with every spatial augmentation to preserve annotation validity.

3.4 RF-DETR Architecture

RF-DETR comprises four principal components:

3.4.1 Convolutional Backbone

The feature extractor is a ResNet-50 network pre-trained on ImageNet. The last fully connected classification layer is discarded [12]. The map dimensions of the maps are 512, 1024, 2048 channels with strides of 8, 16, 32 respectively and are extracted from the 3 residual stages (C3, C4, C5). A 1×1 convolutional projection maps all three maps to an embedding dimension $d = 256$.

3.4.2 Region-Focused Attention Module (RFAM)

RFAM is the novel contribution distinguishing RF-DETR from prior transformer-based detectors. Given a flattened feature sequence $F \in \mathbb{R}^{(N \times d)}$ where N is the number of spatial tokens and $d = 256$, RFAM estimates a scalar saliency weight w_i for each token via a two-layer MLP [13]:

$$s_i = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot f_i + b_1) + b_2)$$

where $W_1 \in \mathbb{R}^{(64 \times 256)}$, $W_2 \in \mathbb{R}^{(1 \times 64)}$, and σ denotes the sigmoid activation. The output saliency map $S = \{s_1, \dots, s_N\}$ modulates the feature sequence element-wise before it enters the transformer encoder: $F' = S \odot F$. Tokens with low saliency scores (corresponding to uniform background patches) are effectively down-weighted, reducing the attention budget consumed by irrelevant context.

3.4.3 Transformer Encoder-Decoder

The modulated sequence F' is fed into a six-layer transformer encoder with multi-head self-attention (8 heads, $d_k = 32$ per head) [14] and position-wise feed-forward networks of hidden dimension 1024. Sine-cosine fixed positional encodings are added prior to the first encoder layer. The decoder consists of six layers with cross-attention between learned object queries ($Q \in \mathbb{R}^{(100 \times 256)}$) and encoder memory, enabling parallel decoding of up to 100 candidate objects [15].

3.4.4 Detection Head

Each decoded query is passed through two parallel branches: a three-layer FFN producing a $(C+1)$ -dimensional class probability vector ($C = 5$ disease categories plus background), and a three-layer FFN predicting a four-dimensional bounding-box coordinate vector (centre x , centre y , width, height) normalised to $[0, 1]$. Training is supervised by the

Hungarian matching loss combining focal cross-entropy for classification and a weighted sum of L1 and generalised IoU (GIoU) losses for localisation [16].

3.5 Training Strategy

The model was implemented in PyTorch 2.1 and trained on a single NVIDIA A100-SXM4-80GB GPU. The following hyperparameters were used [17]:

Table 2. Training Hyperparameters

Hyperparameter	Value
Optimiser	AdamW
Initial learning rate (backbone)	1×10^{-5}
Initial learning rate (transformer & head)	1×10^{-4}
LR scheduler	Cosine annealing with warm restarts ($T_0 = 10$)
Weight decay	1×10^{-4}
Dropout (transformer layers)	0.1
Batch size	8
Training epochs	100 (with early stopping, patience = 15)
Gradient clipping (max norm)	0.1
Backbone pre-training	ImageNet-1k
Classification loss weight (λ_{cls})	1.0
L1 box loss weight (λ_{L1})	5.0
GIoU loss weight (λ_{GIoU})	2.0

Overfitting was mitigated through three complementary strategies. First, the extensive augmentation pipeline described in Section 3.3 increases effective dataset diversity. Second, dropout (rate = 0.1) is applied after each transformer sub-layer. Third, L2 regularisation via weight decay (1×10^{-4}) penalises large parameter magnitudes. The backbone was frozen for the first 10 epochs to allow the transformer head to warm up before joint fine-tuning commenced. Early stopping was performed based on the lack of improvement in validation mAP@0.5 after 15 epochs [18].

3.6 Evaluation Metrics

The performance of a model is evaluated by means of classification and detection specific metrics:

- Accuracy: fraction of correctly classified image-level predictions.

4. Outcomes and Analysis

4.1 Quantitative Performance

Table 3 shows the performance of RF-DETR and all of baseline models on the held-out test set (630 images). All values are the mean of five independent runs (standard deviations are given in parentheses).

Table 3. Mean $\pm$ SD (over 5 runs) of comparative performance on Test Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Mean Average Precision at IoU Threshold 0.5 (%)	Mean Average Precision at IoU Thresholds from 0.5 to 0.95 (%)
CNN	85.1 (0.8)	83.4 (0.9)	82.0 (1.1)	82.7 (0.9)	78.3 (1.2)	61.4 (1.4)
ResNet-50	89.2 (0.6)	88.1 (0.7)	87.0 (0.8)	87.5 (0.7)	84.2 (0.9)	68.7 (1.1)
ViT-B/16	91.3 (0.5)	90.2 (0.6)	89.1 (0.7)	89.6 (0.6)	87.5 (0.8)	72.3 (1.0)
YOLOv11-m	92.4 (0.4)	91.6 (0.5)	90.8 (0.6)	91.2 (0.5)	90.1 (0.7)	76.8 (0.9)
EfficientDet-D3	91.8 (0.5)	90.9 (0.6)	90.0 (0.7)	90.4 (0.6)	89.3 (0.8)	75.2 (1.0)
RF-DETR (Ours)	95.1 (0.3)	94.2 (0.4)	93.0 (0.5)	93.6 (0.4)	93.1 (0.6)	81.4 (0.8)

The test set results show that the overall mean IoU for RF-DETR is 87.6%, which is higher than YOLOv11-m (81.3%) and YOLOv11 (79.8%). EfficientDet-D3, which proves its high precision in the bounding-box area.Per-class mAP@0.5 values are: Red Rot 94.2%, Rust 93.8%, Leaf Scald 92.6%, Yellow Leaf 91.9%, Healthy 93.1%.

4.2 Model Component Evaluation

To understand the contribution of RFAM, four different variants of the model were tested on the validation set:

Table 4. Model Component Evaluation Results on the Validation Set

Model Variant	Configuration	mAP@0.5 (%)	Mean IoU (%)
A: Backbone only	ResNet-50 + head	84.1	78.2
B: Backbone + Transformer	ResNet-50 + DETR	88.7	82.9
C: Backbone + RFAM + Head	No transformer encoder	89.4	83.6
D: Full RF-DETR	All components	93.1	87.6

The difference between variants B and D demonstrates the benefit (+4.4% in mAP@0.5 and +4.7% in mean IoU) of region-focused attention, which highlights the fact that this is the most important factor for better localisation performance. The isolated effect of the transformer (B vs. A) improves mAP by 7.6% at 0.5. Without the full transformer encoder (C vs. A), the mAP increases by 7.6%. A) produces a +5.3% improvement, demonstrating that the saliency filtering of RFAM works even without the entire transformer stack. The synergistic benefits of the combined architecture (D) is greater than the sum of its individual parts.

4.3 Overfitting Analysis

The training and validation loss curves were observed for every 100 epochs. The difference in training mAP@0.5 (94.7%) and validation mAP@0.5 (93.4%) is 1.3 percentage points, which is very low and shows minimal overfitting. A convergence of the training and validation loss curves, between epochs 60 and 80, and the subsequent leveling off in the validation loss before early stopping activation (at epoch 87) indicate that the loss function is sufficiently regularised. This five-run standard deviation of 0.6% on mAP@0.5 further validates the stability of the models across different data splits.

5. Existing Challenges and Future Research

However, there are a number of limitations to note. The dataset, while carefully curated, was collected from a geographically limited growing region and two growing seasons. Generalisation to other agro-climatic zones and cultivar types requires broader sampling. The inference latency of RF-DETR on a mobile CPU is approximately 340 ms per image,

which is insufficient for real-time drone-mounted processing and must be addressed through model compression techniques such as knowledge distillation or quantisation-aware training.

Future work will pursue: (i) expansion of the dataset to include multi-regional, multi-seasonal imagery; (ii) development of a lightweight RF-DETR variant targeting edge deployment on NVIDIA Jetson Nano; (iii) integration of multispectral imagery to detect sub-visual early-stage infections not visible in RGB; and (iv) extension of the annotation protocol to support instance segmentation, enabling pixel-level disease area quantification.

6. Conclusion

The paper discussed RF-DETR, a Region Focus Detection Transformer which was designed for detecting sugarcane leaf diseases. The model comprised of a ResNet-50 convolutional backbone network, a learnable Region Focus Attention Module which applies spatial weighting to discriminative lesion tokens, and a transformer encoder-decoder which models global contextual relationships. Upon training on an annotated 4,200 image dataset consisting of stratified splits and validated using statistical tests over five runs, RF-DETR attained mAP@0.5 of 93.1%, mAP@0.5:0.95 of 81.4%, and mean IoU of 87.6%. It outperformed other baseline models including CNN, ResNet-50, ViT, YOLOv11, and EfficientDet. Experiments performed demonstrated that the Region-Focus Attention Module was responsible for the improvement of the localization abilities of the model, improving mAP@0.5 score by +4.4% over the transformer model without the region focus attention module.

References

1. Bischoff, E., Vladeva, S., Baraliakos, X., & Kirilov, N. (2026). AI-Powered Lateral DEXA Morphometry for Integrated Evaluation of Thoracic Kyphosis and Bone Density Assessment in Patients with Axial Spondyloarthritis. *Life*, 16(1), 162. <https://doi.org/10.3390/life16010162>
2. Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2023. (2024). Springer Nature Singapore. <https://doi.org/10.1007/978-3-031-72114-4>
3. B. G., N., & Kannadhasan, S. (Eds.). (2026). Information and Communication Systems (1st ed.). CRC Press. <https://doi.org/10.1201/9781003650201>
4. Kumar, P., Luo, S., & Shaukat, K. (2023). A comprehensive review of deep learning approaches for animal detection on video data. *International Journal of Advanced Computer Science and Applications*, 14(11). [https://thesai.org/Downloads/Volume14No11/Paper_144-A_Comprehensive_Review_of_Deep_Learning_Approaches .pdf](https://thesai.org/Downloads/Volume14No11/Paper_144-A_Comprehensive_Review_of_Deep_Learning_Approaches.pdf)

5. Ferro, P., Vemanaboina, H., & Prakash, C. (Eds.). (2026). *Computational Techniques and Smart Manufacturing* (1st ed.). CRC Press. <https://doi.org/10.1201/9781003679622>
6. Patil, A. R., & Bansode, R. (2026). EDGEAWARE: Distributed threat detection and monitoring system for smart cities. *Interdisciplinary Journal of AI, Machine Learning & Data Science (IJAIMLDS)*, 1(2), Article e005. <https://doi.org/10.66261/falcep06>
7. Ignasius, D., Dewi, I. N., Norman, M. B. C., & Sani, R. R. (2025). Medical named entity recognition from Indonesian health-news using BiLSTM-CRF with static and contextual embeddings. *Journal of Applied Informatics and Computing (JAIC)*, 9(6), 2974–2985. <http://jurnal.polibatam.ac.id/index.php/JAIC>
8. Kovtun, V. Adaptive information-constrained mapping for feature compression in edge AI and federated systems. *Sci Rep* **15**, 30915 (2025). <https://doi.org/10.1038/s41598-025-16604-2>
9. Akkawi, A. N., et al. (2024). Vision-based sensor system for offshore crane pose estimation (Master's thesis). Norwegian University of Science and Technology, Faculty of Engineering, Department of Mechanical and Industrial Engineering. <http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/16731/P30208.pdf?isAllowed=y&sequence=1>
10. Parveen, S., Sanjana, S., Sakshi, S., & Kumar, E. (2026). Early sepsis prediction in ICU patients using deep learning: A comparative modeling study of Transformer, TAGF, CNN-BiLSTM, and XGBoost. Research Square. <https://doi.org/10.21203/rs.3.rs-9499930/v1>
11. Kumar, P., Luo, S., & Shaukat, K. (2023). A comprehensive review of deep learning approaches for animal detection on video data. *International Journal of Advanced Computer Science and Applications*, 14(11). <https://doi.org/10.14569/IJACSA.2023.01411144>
12. Lin, Z., Wang, L., He, R., Zha, H., Zhou, J., & Liu, C.-L. (Eds.). (2019). *Pattern recognition and computer vision: Second Chinese Conference, PRCV 2019, Xi'an, China, November 8–11, 2019, proceedings, part I (Lecture Notes in Computer Science, Vol. 11857)*. Springer. <https://doi.org/10.1007/978-3-030-31654-9>
13. Wang, C., He, N., Wang, B. (2025). Chareption: Change-Aware Adaption Empowers Large Language Model for Effective Remote Sensing Image Change Captioning. In: Lin, Z., *et al.* *Pattern Recognition and Computer Vision. PRCV 2024. Lecture Notes in Computer Science*, vol 15043. Springer, Singapore. https://doi.org/10.1007/978-981-97-8493-6_24
14. Bhattarai, A., Jin, J., & Kambhamettu, C. (2024). Analyzing adjacent B-scans to localize sickle cell retinopathy in OCTs. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Springer. https://papers.miccai.org/miccai-2024/paper/3665_paper.pdf
15. ProQuest Dissertations & Theses Global. <https://www.proquest.com/docview/3319232943>

16. Nguyen, T. (2025). Improve underwater object detection through YOLOv12 architecture and physics-informed augmentation. arXiv. <https://doi.org/10.48550/arXiv.2506.23505>
17. Kumar, P., Luo, S., & Shaukat, K. (2023). A comprehensive review of deep learning approaches for animal detection on video data. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 14(11). <https://doi.org/10.14569/IJACSA.2023.01411144>
18. Nguyen, T. (2025). Improve underwater object detection through YOLOv12 architecture and physics-informed augmentation. arXiv. <https://doi.org/10.48550/arXiv.2506.23505>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

