Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Explainable Hybrid CNN–Transformer Framework for Crop Classification and Biotic Stress Detection under Real-World Field Variability

Muhammad Majid
Department of Computer Sciences, University of Agriculture, Faisalabad-38040, Pakistan

Shahzad Nazir*
Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan
Email: shahzad.nazir@bgnu.edu.pk

ABSTRACT

In agricultural field applications, reliable image analysis is still challenging due to the variability of the crop appearance with illumination, background clutter, camera viewing angle, growth stages, partial occlusion and visible pest and disease symptoms. This study proposes a hybrid CNN–Transformer architecture to classify crops and analyze biotic stresses in such heterogeneous settings, with explainability. The framework relies on the ResNet-50 convolutional backbone to learn the local morphology and texture, translates high-level feature maps into tokens, and a Transformer encoder to capture the long range contextual dependencies. Fused local-global features are served to task-specific prediction heads, and visual explanations are given by Grad-CAM and Transformer attention maps. Experiments are performed on a well curated dataset of 30,204 RGB images from 15 different crop categories with standardized input size of 224×224, augmentation used only during training, stratified partitioning and imbalance-aware optimization. The obtained results of the implemented model were: the accuracy for crop classification was 97.31%, the weighted precision was 0.970, the weighted recall was 0.9731, the weighted F1-score was 0.971, the macro F1-score was 0.934 and the balanced accuracy was 0.940. The accuracy of multi-seed testing was 97.31±0.07%, which was relatively stable. The proposed model also achieved the best performance among the following models under the same evaluation protocol: VGG16, MobileNetV2, ResNet-50, DenseNet121, and EfficientNetB0. The findings are in support of the use of hybrid representation learning, with a local and a global representation, for robust and interpretable agricultural decision support. When exported from the implemented run, separate stress-head metrics should be reported.

Keywords: Crop Classification; Biotic Stress; CNN–Transformer; Explainable AI; Grad-CAM; Precision Agriculture; Field Variability; Class Imbalance.

Introduction

The use of computer vision and deep learning in agriculture is growing rapidly, aiming to automate crop monitoring, disease surveillance, pheno typing and decision

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

support. The popularity of image-based methods is due to the benefits of using a RGB camera, which is low cost, easily expandable, and supported by many platforms including ground platforms, smartphones, and unmanned aerial systems. A number of the early studies in deep learning for agriculture showed that convolutional neural networks (CNNs) can be trained directly on pixels, while also providing a significant improvement over manual feature engineering pipelines (Kamilaris & Prenafeta-Boldú, 2018; Mohanty et al., 2016; Ferentinos, 2018; Too et al., 2019). The benefits of transfer learning and end-to-end visual representation learning for the application of practical agriculture recognition have been further demonstrated in subsequent works on cassava (Ramcharan et al., 2017), tomato (Fuentes et al., 2017), maize (Barbedo, 2019), and cotton.

Even with the advances, the difficulty of real-world images of crops is still far greater than controlled leaf images. In addition to these factors, soil, weeds, neighboring plants, non-uniform lighting, viewpoint changes, growth stage variations, insect and disease effects can impact the colour, texture, and geometry of leaves in field scenes (Zheng et al., 2024; Ngugi et al., 2024; Shahid et al., 2024; Barbedo, 2019). All of this adds to a high intra-class variance and can lead to artificial correlations between crop labels and the context. Long-tailed class distributions exacerbate this issue, as a model might be accurate overall, but fail to perform well on minority crops. Furthermore, the accuracy of black-box predictions is hard to believe if the model's classification is not based on the morphology and stress symptoms of the crop, but rather contains irrelevant background regions (Selvaraju et al., 2017; Lin et al., 2017; Cui et al., 2019).

The literature has shifted from traditional CNN backbones to attention mechanisms and Transformer-based models. Residual networks boost optimization of deep CNNs (He et al., 2016), DenseNet, MobileNetV2 and EfficientNet offer various degrees of feature reuse, efficiency, and accuracy (Sandler et al., 2018; Huang et al., 2017; Tan & Le, 2019). A series of works have emerged in a short period, such as the introduction of patch-based self-attention for global image modeling by Vision Transformers (ViT) (Dosovitskiy et al., 2021), and then DeiT and Swin Transformer which improved data efficiency and hierarchical representation learning (Touvron et al., 2021; Liu et al., 2021). Reedha et al. demonstrated that Transformer models can be used for crop and weed classification in agricultural imagery using high-resolution UAV images (Reedha et al., 2022). Meanwhile, attention modules like squeeze-and-excitation (SE) and CBAM showed that feature recalibration can diminish irrelevant answers (Hu et al., 2018; Woo et al., 2018); and Grad-CAM has turned into a broadly used way to clarify visual answers (Selvaraju et al., 2017).

Inspired by these advances, this work introduces a hybrid CNN–Transformer model that integrates a complementary local and global representation and is explainable. Pretrained ResNet-50 extracts local texture, edge, venation, morphology and lesion-related features, followed by a projection of high level feature maps into tokens, which are modeled using multi-head self-attention. CNN and Transformer descriptors are then fused for the classification of crops and biotic-stress-aware prediction. The experimental protocol maintains the same set of images and image processing procedure as specified before: 30,204 RGB images, 15 crop classes, 224×224 input resolution, 70/15/15 partitioning (70% training, 15% validation, 15% test), training-

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

only rotation, horizontal flipping and brightness/contrast augmentation, and class-weighted optimization. Grad-CAM and Transformer attention visualization provide explanations.

The experimental results demonstrate that the proposed model achieves 97.31% accuracy in crop-classification, which is 4.31 percentage points higher than the baseline model ResNet-50 with 93.00%. Weighted precision, recall, and F1-score are 0.970, 0.9731, and 0.971, respectively, while macro precision, recall, and F1-score are 0.930, 0.940, and 0.934, and balanced accuracy is 0.940. For each of the three random seeds, the accuracy is 97.31±0.07%. Under the same protocol, the proposed approach also exceeds VGG16 (91.84%), MobileNetV2 (92.76%), DenseNet121 (95.42%), and EfficientNetB0 (96.18%). The results show that a convolutional inductive bias combined with global self-attention can enhance the robustness of crop recognition for heterogeneous field conditions.

Literature Review

Deep Learning for Crop and Plant Image Classification

The paper revolutionized agricultural image analysis by introducing deep learning to replace hand-crafted color, shape, and texture descriptors by learned hierarchical features. The PlantVillage dataset was used to prove the feasibility of CNN-based disease classification (Mohanty et al., 2016) and several deep architectures were tested on a large multi-crop disease dataset, and good diagnostic results were observed (Ferentinos, 2018). Too et al. also demonstrated that fine-tuned pretrained CNNs can achieve good performance in plant disease recognition, especially when the number of plants in the specific dataset is significantly smaller than the number of images in the generic computer-vision datasets (Too et al., 2019). In other applications, the CNN was also observed to learn discriminative patterns of disease symptoms that were challenging to encode using conventional image processing, such as cassava and tomato disease recognition (Ramcharan et al., 2017; Fuentes et al., 2017).

But high scores for clean leaf images don't necessarily carry over to real-world images. Barbedo demonstrated that lesion size, symptom complexity, background variability, and image diversity affect the performance of deep models (Barbedo, 2019). Similarly, when it comes to transfer-learning architectures, Moin et al. (2022) reported variations between crop types and disease categories in their investigation of crop-diseases in Bangladesh. However, more recent studies with a field focus have focused on the importance of locally gathered data, augmentation, and ensemble learning. There is a need to assess models beyond image collections like those in a laboratory setting, as in the example of Shahid et al., who applied field-acquired cotton imagery and transfer learning to determine healthy and unhealthy plants (2024).

Model architecture also affects the accuracy-efficiency trade-off. VGG16 has deep stacks of small convolutional filters (Simonyan & Zisserman, 2015), MobileNetV2 has inverted residuals and linear bottlenecks for efficient inference (Sandler et al., 2018), DenseNet has dense connectivity for feature reuse (Huang et al., 2017) and EfficientNet jointly scales depth, width and resolution (Tan & Le, 2019). These architectures are still significant standards for crop and disease identification and will

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

be used as a part of the controlled comparison in this study.

Attention and CNN–Transformer Representation Learning

CNNs are good at local pattern extraction but explicit modelling of distant relationships may be constrained by their inductive bias on local neighbourhoods. To overcome this limitation, Vision Transformers have been proposed, which process an image as a series of tokens and then use self-attention over the sequence of tokens (Dosovitskiy et al., 2021). DeiT minimized the data and computational requirements for training ViT early (Touvron et al., 2021), and Swin Transformer added hierarchical shifted-window attention for multi-scale vision tasks (Liu et al., 2021). These advances led to the hybrid designs where convolution is used for local structure and Transformer blocks are used for global context.

One important aspect of this complementarity is that a prediction may rely on both fine texture of leaves and general architecture of the plants. Agricultural scenes are particularly well suited for this complementarity, due to the need for a prediction to rely upon both finer texture of leaves and broader plant structure simultaneously. However, Reedha et al. (2022) showed the potential of the self-attention mechanism in agriculture by applying it to weed and crop recognition in high-resolution Unmanned Aerial Vehicle (UAV) images. At the same time, light-weight attention mechanisms like squeeze-and-excitation and CBAM demonstrated selective attention to informative channels or spatial regions can enhance discriminative representations without the need to replace the convolutional backbone (Hu et al., 2018; Woo et al., 2018). These types of attention mechanisms are especially valuable when there is a mixture of soil, weeds or background vegetation present that may be relevant to the crop.

This local-global rationale is not to replace convolution but rather used as the proposed framework. ResNet-50 is kept as an effective local feature extractor (He et al., 2016), with high-level convolutional features being mapped to Transformer tokens. This design mitigates the need to learn low level visual structure only by self-attention while retaining the ability to make long-range connections between image regions. The fusion stage then delivers a common representation that can be used to support decisions and crop identity when under stress.

Explainability, Field Variability, and Imbalance-Aware Learning

In the context of agriculture AI, the key to interpretability is that a right class label is not enough if the AI reaches it for the wrong visual reason. Grad-CAM is an approach that works by analyzing the CNN's feature map and highlighting the regions that are most important for the predictions, and can be used to uncover whether predictions are based on leaves, lesions, stems or on irrelevant background context (Selvaraju et al., 2017). Activation maps are becoming increasingly popular in explainable plant-disease studies to examine the behaviour of the system, especially if it is to be deployed in the field. While attention maps can provide complementary information about interactions between regions of the image when they are tokenized, attention maps should be considered as evidence and not as a full causal explanation.

Experimental validity is also dependent upon field variability, as is augmentation and leakage control. Another common technique to enhance the visual diversity and

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

decrease the overfitting is data augmentation when the acquisition condition changes (Shorten & Khoshgoftaar, 2019). Meanwhile, splitting before the augmentation is necessary since the augmented versions of the same image should not be present in both the training and testing partitions. In addition, long-tailed recognition needs extra attention – easy examples can be underrepresented by focal loss (Lin et al., 2017) and learning adjusted by class-balanced weighting methods (Cui et al., 2019) can be explicitly adjusted. These set ideas drive the idea of the class-weighted loss and macro-level evaluation in the present study.

The past few years' agricultural literature also has focused on the need for generalization in complex environments. Self-attention has been integrated into lightweight drone image-based crop segmentation (Zheng et al., 2024); persistent gaps in crop disease diagnosis under variable data conditions were identified (Ngugi et al., 2024); and transfer learning-based crop disease recognition schemes have been shown to perform well with varying degrees of strength, depending on the specific data set (Paymode & Malode, 2022; Atila et al., 2021). Field experiments, crop cover photos, UAV images, and insect recognition also indicates that agricultural models should be evaluated with variation of background, class similarity, and acquisition setting (Haque et al., 2022; Yordanov et al., 2023; Silva et al., 2024; Ai et al., 2020; Kasinathan et al., 2021).

Materials and Methods

This section explains the experimental methodology implemented and evaluated the proposed framework. It starts with dataset curation and stratified partitioning, then it moves on to resizing, intensity scaling, and augmenting for training only. First, ResNet-50 CNN extracts local features, then, tokenize the local features and send them to a Transformer encoder to learn global dependencies, followed by feature fusion and task-specific prediction. The training protocol employs imbalance aware optimization and fixed evaluation splits, and Grad-CAM and Transformer attention offer explanations to the model. The overall process is summarized in Figure 1.

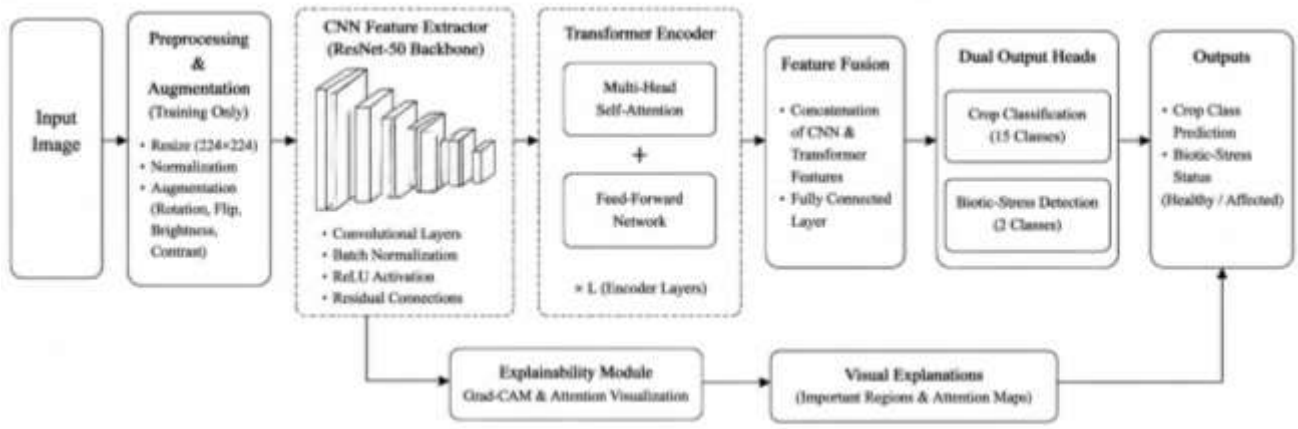

Figure 1. Overall methodology of the proposed explainable hybrid CNN–Transformer framework.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Dataset Description

The experiments utilize the same selected agricultural RGB image set that was used in the previous crop-classification experiment. The final dataset consists of 30,204 images across 15 crop classes, which is kept as is – without changing the class definitions. The collection contains diverse agricultural imagery that includes differences in direct light, shadow, partial overexposure, background soil and plant textures, camera angle, crop size, stage of development, dust, moisture, motion blur, some areas partially covered, and a variety of pest and disease symptoms. This diversity makes it more representative of conditions that are likely to be met in deployment than controlled-background collections of leaves. The underlying public resources for crop diseases are associated with the collection described by Moin et al. (Moin et al., 2022) and the curated project release has been archived at Zenodo (Majid et al., 2026).

The data set is highly skewed with the skewness being observed in Table 1. Maize-I consists of 14,632 images while Maize-II consists of only 200 images. There are also several categories of pulses with between 400 and 1,200 samples, and of oilseeds with between 200 and 600 samples, and rice and wheat are also relatively large categories. This imbalance can induce over-representing the dominant class, and this is the reason why the stratified splitting, class-weighted training and reporting of macro as well as prevalence-weighted metrics are encouraged (Lin et al., 2017; Cui et al., 2019).

It is important that the visible biotic stress cues are not excluded, as they are part of the field appearance. The main supervision target for crop classification is the original crop identity. If verified stress annotations are available, the proposed framework also introduces an output pathway for healthy-versus-stressed prediction, based on stress. Since the provided experimental result summary is quantitative for the crop-classification output, the Results section is limited to that output, and does not create separate stress-head scores.

Table 1. Class distribution of the curated agricultural image dataset used in this study.

Dataset characteristic	**Value**
Total images	30,204
Number of crop classes	15
Largest class	Maize-I — 14,632
Smallest class	Maize-II — 200
Class-size range	200–14,632
Major additional classes	Wheat — 3,781; Rice — 2,791

The ratio of the largest to smallest categories is more than 70:1 as outlined in Table 1. This property renders overall accuracy as a standalone success measure to be insufficient and necessitates the addition of balanced accuracy and macro-averaged

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

precision, recall and F1-score metrics to the weighted summary metrics for success.

Data Preprocessing

All RGB images are downsized to 224×224×3 to ensure a uniform input tensor to the ResNet-50 backbone and the downstream Transformer pathway. Intensities of the pixels are scaled during optimization to guarantee numerical stability. It is deliberately restricted to transformations that are applied to the previous experimental pipeline: random rotation, horizontal flipping, brightness/contrast adjustment. These transformations are representative of camera movement and lighting that mimic biologically realistic modifications to the structure of the crop, but not extreme ones. Data augmentation is aligned with previous guidelines for enhancing generalization where images are not diverse or acquisition is not ideal (Shorten & Khoshgoftaar, 2019).

The entire data set is partitioned prior to augmentation, with 70% used for training, 15% for validation and 15% for an independent data set for testing. Only the training subset is augmented, no validation or test images are augmented. To minimise the risk of leakage, duplicate and near duplicate images are removed during curation. The main split is performed using a fixed primary seed (42), with repeated training runs performed using a primary seed of 42, 123 and 999 to assess stochastic stability. This protocol maintains the class proportions within subsets and guarantees that the test statistics are computed for unseen original images and not augment derivatives.

Proposed Hybrid CNN–Transformer Model

CNN Feature Extraction

It is based on the ImageNet-pretrained ResNet-50 backbone (He et al., 2016) in the local representation branch. Residual learning can aid in optimizing deep convolutional layers while maintaining the multi-level feature reuse. An input image I is fed into the CNN, which outputs a high-level feature tensor F_C, whose channels include local crop morphology, edges, textures, venation patterns and appearance cues related to stress.

$$F_C = f_C NN(I; \theta_C) \quad (1)$$

Before the original classification layer, the final convolutional tensor is saved. This ensures the spatial information needed for tokenization and Grad-CAM explanation. Transfer learning is applied to ensure that the generic low-level visual filters are adapted to agricultural images instead of being learned from the long-tailed dataset.

Tokenization and Transformer Encoder

The convolutional feature map is transformed into a series of spatial tokens. A learnable linear projection projects each token to the Transformer embedding dimension, to allow the encoder to capture spatial relationships between image regions. Let Flatten(F_C) be the sequence obtained from CNN's feature map, then the first Transformer representation is:

$$Z^0 = Proj\big(Flatten(F_C)\big) + E_pos \quad (2)$$

The sequence of tokens passes through multiple blocks of the Transformer's encoder, consisting of stacked layers of Transformer blocks with residual connections, multi-

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

head self-attention, and layer normalization. In addition to the local receptive field of convolution, self-attention enables a crop region to directly interact with faraway regions of the same image. Given matrices Q, K and V, scaled dot-product attention is calculated as follows:

$$Attention(Q, K, V) = softmax\left(QK^{\mathrm{T}}/\sqrt{d_k}\right)V \quad (3)$$

This is especially helpful when discriminative evidence is spatially distributed, such as when crop identity is a function of both leaf form and canopy context, or where stress cues are found in multiple non-contiguous regions. Furthermore, the CNN front end alleviates the requirement of pure ViTs to learn low-level edge and texture representations from scratch, which is beneficial in agricultural datasets where the number of images is smaller than the large corpora used in the pre-training of pure ViTs (Dosovitskiy et al., 2021; Touvron et al., 2021; Liu et al., 2021; Reedha et al., 2022).

Feature Fusion, Prediction Heads, and Explainability

Global average pool of CNN feature tensor and aggregation of the final Transformer representation before combining with local descriptors. Concatenation is used to keep complementary information from both branches:

$$F_fuse = Concat(GAP(F_C), z_T) \quad (4)$$

Fused vector is fed into the task-specific fully connected layers. The crop head applied a 15-class softmax output, and the stress-aware head was developed for the binary classification task of healthy versus stressed classification when verified annotations are available for the corresponding crop head. The training objective is the sum of the losses due to weights, λ_c and λ_s:

$$L_total = \lambda_c L_crop + \lambda_s L_stress \quad (5)$$

The model is rendered explainable at two levels. To identify which local image regions are most important for a prediction, the Grad-CAM is computed on the last convolutional feature maps (Selvaraju et al., 2017). Transformer attention maps are examined to gain a visualization of general token-to-token relationships. These explanations, combined, can be used to decide whether predictions are based on crop structures and visible stress patterns, or soil, sky, neighbouring plants, or other spurious background cues.

Training and Optimization

The framework is trained using the stochastic gradient descent (SGD) with momentum method, as done previously for ResNet training. Class-weighted cross-entropy is used for the crop-classification head because of the large differences in class frequency. An inverse frequency weight is computed for each crop class k, which is w_k = N/K n_k, where N is the number of training samples, K is number of crop classes and n_k is number of training samples in class k. The weighting escalates the penalty for minority class errors, and is conceptually similar to imbalance aware learning strategies (Lin et al., 2017; Cui et al., 2019).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 2. Training configuration used for the proposed model and controlled baselines.

Hyperparameter	**Setting**
Input size	224×224×3
Optimizer	SGD
Momentum	0.9
Initial learning rate	0.01
Learning-rate schedule	Step decay; γ=0.1 every 10 epochs
Weight decay	1×10^{-4}
Batch size	32
Epochs	50
Crop loss	Class-weighted cross-entropy
Primary seed	42
Robustness seeds	42, 123, 999

The same optimizer, batch size, learning-rate policy, input size, augmentation policy and data partitions are used in all of the controlled baseline experiments (as seen in Table 2). This is crucial as differences in performance should be attributed to the model architecture, not the training recipe.

Explainability Analysis

For representative samples of correctly classified and misclassified samples, grad-CAM heat maps are produced. The leaves, stems and the contour of the canopy are important features for crop classification, while background items are not. The areas in which high response is expected are lesions, discoloration, feeding damage, necrosis, and other symptom expression would be considered high response areas when making stress-aware predictions. In addition to Grad-CAM, transformer attention maps are also examined to check long-range dependencies. The two explanation mechanisms are complementary: Grad-CAM shows convolution evidence that is class-sensitive, and attention visualization shows how globally-separated regions interact during representation learning (Selvaraju et al., 2017).

Explainability as a model-audit mechanism is used primarily, and not as a replacement to quantitative performance. If the correct prediction always depends on irrelevant background areas, the model could be using the bias of the dataset and not be applicable to other farms or acquisition devices. On the other hand, focusing on crop structures and symptom regions offers more evidence of the learning of agriculturally meaningful features by the network. This is particularly important for the field images where background-clutter correlations can be high (Shahid et al.,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

2024; Zheng et al., 2024; Ngugi et al., 2024).

Ablation and Comparative Experimental Protocol

Ablation analysis is employed to assess whether the complete proposed configuration is worth the residual backbones that exist in the system. All of the ResNet-18, ResNet-34, and ResNet-50 were trained with the same split and optimization protocol. Finally, the proposed run is compared with a ResNet-50 baseline to measure the gain from using the full hybrid representation. Additionally, the models VGG16 (Simonyan & Zisserman, 2015), MobileNetV2 (Sandler et al., 2018), DenseNet121 (Huang et al., 2017), or EfficientNetB0 (Tan & Le, 2019) are reproduced with the same preprocessing and evaluation procedure. This is a same-data, same-protocol design that is common when comparing the accuracy scores of different datasets, class definitions, or train-test partitions.

In order to assess the sensitivity to stochastic initialization and mini-batch ordering, repeated training is conducted with seeds 42, 123 and 999 for the proposed model.The proposed model is trained with seed 42, 123, and 999 to assess the sensitivity of the model to the stochastic initialization and mini-batch ordering. The accuracy and weighted summary metric are provided as individual seed results and as mean ± standard deviation. The independent test set is not used for model selection, but rather the 15% test set.

Evaluation Metrics

The performance of the model is evaluated through accuracy, precision, recall, F1-score, macro average, weighted average, confusion matrix analysis and balanced accuracy. Weighted metrics take into consideration the number of test samples from each class, while macro metrics give equal weight to each crop category and thus are important for a long-tailed dataset. Precision, Recall and F1-score are calculated for a particular class using true positives (TP), false positives (FP) and false negatives (FN) as follows:

$$Precision = TP/(TP + FP) \quad (6)$$

$$Recall = TP/(TP + FN) \quad (7)$$

$$F1 = 2 \times Precision \times Recall/(Precision + Recall) \quad (8)$$

The overall accuracy is the ratio of correctly classified test samples to the total number of test samples. Balanced accuracy is the average of the K crop categories' recalls, avoiding the dominant class from having a disproportionate influence on the summary score:

$$BalancedAccuracy = (1/K)\Sigma_k Recall_k \quad (9)$$

The weighted, macro and class-balanced metrics are based on the standard metrics for imbalanced visual recognition (Lin et al., 2017; Cui et al., 2019). All the same metric definitions are used for all the baseline and ablation setups in the current experiments.

Results and Discussion

Effect of Preprocessing and Experimental Control

The pipeline set up during the preprocessing phase ensured uniformity of the heterogeneous field images prior to optimization of the models. The pretrained ResNet-50 network was found to be incompatible with a variety of sample sizes, so

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

all samples were resized to 224×224×3 to ensure compatibility.The pretrained ResNet-50 network was not compatible with all sample sizes, so all samples were resized to 224×224×3 to ensure compatibility and a fixed spatial basis for Transformer tokenization. Intensity scaling minimized the variance of numerical values and rotation, horizontal flipping, and brightness/contrast variations introduced likely changes in illumination and orientation within the model. The operations are especially important for the present data set as the original images include direct sun light, shadows, varying viewing angles, cluttered backgrounds, differences in growth stages, and biotic-stress indicators.

One of the most important design decisions made was to split the data set prior to augmentation. Therefore, modified copies of a training image cannot be included in the validation or test sets. Further, duplicate and near duplicate removal reduced the chance that visually similar samples were on either side of the train-test boundary. This is an experimental control, because even a small leak, in particular in large image collections with multiple similar observations in the pictures, can lead to inflated image-classification results.

This run summary is provided without a control run to verify accuracy gains obtained through preprocessing alone, so a pure gain in accuracy due to preprocessing without fabrication cannot be directly related to this run summary. What it does say is that its effect is observed in the stability of the full training pipeline, as the proposed model was within 97.24–97.38% accuracy for three independent seeds (Section 4.2). This relative low variance suggests that the standardized data pipeline and fixed split protocol was a basis for reproducible comparison of the architectures.

The preprocessing stage can be thought of as an enabler for the experimental pipeline and should not be interpreted as a source of improvement on its own. This run summary is not provided with a raw-input control, and therefore no independent gain in preprocessing is claimed. However, the low variance among the three repeated runs is indirect verification that the procedure used for the fixed resizing, scaling, augmentation, stratification and leakage control produced a stable optimization setting. As shown later in Table 4 and Figure 3, the full pipeline only deviated by 0.14 percentage points from its lowest and highest test accuracy.

Performance of the Proposed Model

The performance of the proposed hybrid CNN-Transformer model was tested using a held out 15% of data, with the test accuracy being 97.31%. Both prevalence-weighted and class-balanced measures are provided in Table 3. The weighted precision was 0.970, the weighted recall was 0.9731, and the weighted F1-score was 0.971. The macro precision, macro recall and macro F1 score were 0.930, 0.940 and 0.934, respectively, and balanced accuracy was 0.940. The difference between the weighted metric and the macro is to be expected as Maize-I makes up the bulk of the data set, but macro scores are still high suggesting that the model's performance cannot be attributed entirely to the majority class.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 3. Aggregate test-set performance of the proposed model.

Metric	Weighted average	Macro / balanced
Accuracy	0.9731 (97.31%)	—
Precision	0.970	0.930
Recall	0.9731	0.940
F1-score	0.971	0.934
Balanced accuracy	—	0.940

The major test-set metrics in Table 3 are plotted in Figure 2. Accuracy, weighted precision, weighted recall and weighted F1 score are all above 97% and the balanced accuracy is 94%. This pattern is significant when it comes to a long-tailed problem as a high weighted score with very low balanced accuracy would suggest that the majority classes are suppressing weak minority class recognition. In this case, the 0.940 balanced accuracy is indicative of reasonably uniform sensitivity across classes.

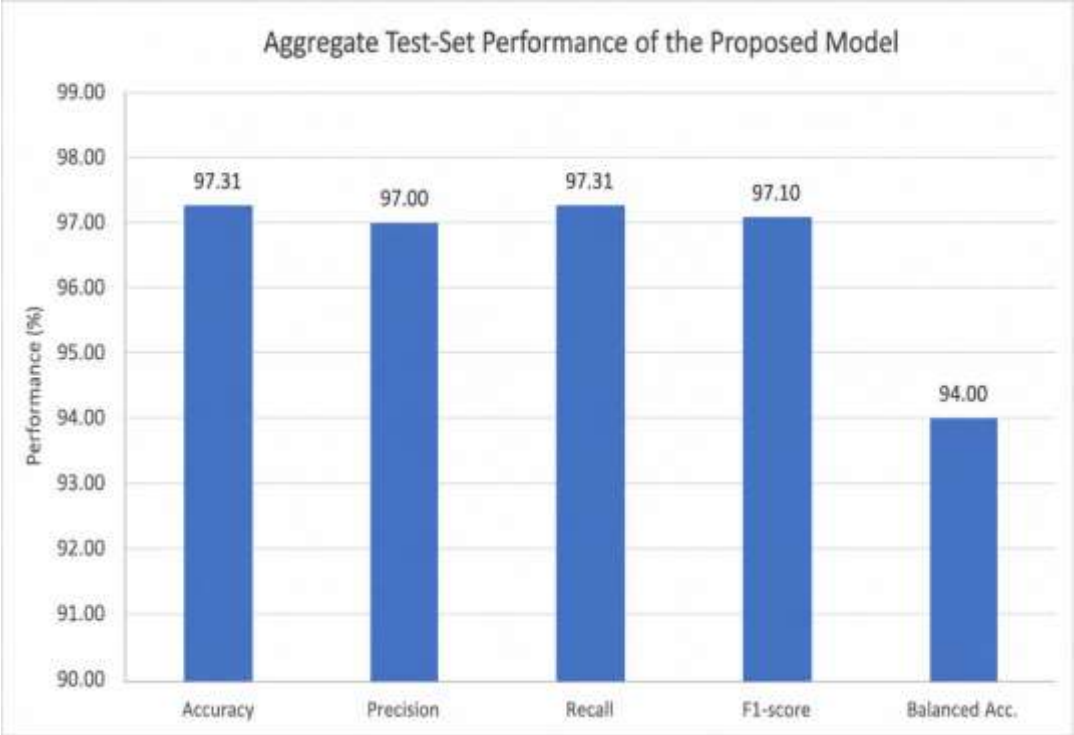

Figure 2. Aggregate test-set metrics achieved by the proposed model.

For assessing stochastic training effects on robustness, the complete model was trained with identical data partitions and hyperparameters for three different training seeds. Table 4 shows accuracies of 97.24%, 97.31%, and 97.38% for seeds 42, 123, and 999, respectively. The mean accuracy was 97.31%, and the standard deviation the mean accuracy was 0.07% points. Weighted precision and F1-score also only fluctuated slightly between runs.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 4. Multi-seed robustness evaluation of the proposed model.

Seed	Accuracy (%)	Weighted precision	Weighted recall	Weighted F1
42	97.24	0.968	0.9724	0.970
123	97.31	0.970	0.9731	0.971
999	97.38	0.972	0.9738	0.973
Mean ± SD	97.31 ± 0.07	0.970 ± 0.002	0.9731 ± 0.0007	0.971 ± 0.002

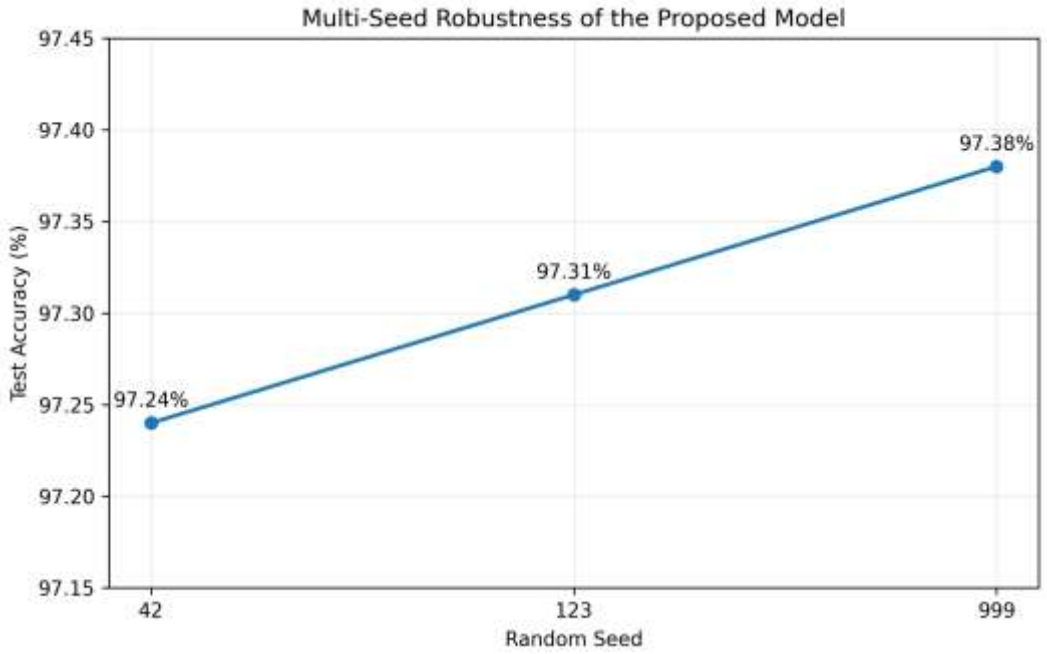

Figure 3. Test accuracy of the proposed model across three random seeds.

The narrow range given in Figure 3 indicates that the result 97.31% is not due to a single lucky initialization of the experiment. This stability enhances the architectural comparisons below. However, it is important to note that the provided experimental summary provides quantitative measurements of the output of crop-classification only. In this paper, the numeric performance of the biotic-stress prediction head is not fabricated, as this is not provided in the supplied run file.

The overall error rate on the held-out set is only 2.69% with the 97.31% test accuracy. The proposed model reduces the observed classification error by ~61.6%, compared to the baseline of 93.00% ResNet-50, which has a classification error rate of 7.00%. This relative error reduction, which gives a complementary perspective to the 4.31 percentage-point gain in accuracy on the residual backbone, shows that the improvement is significant, even in the face of a strong residual backbone. This overall result, however, is understood in conjunction with the macro and balanced measures, due to the fact that the data set is long-tailed.

It is useful to point out the difference between the two kinds of weighted prevalence

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

measures and the class-balanced measures. The weighted precision, the weighted recall, and weighted F1 are by 0.040, 0.0331, and 0.037 respectively more than the macro counterparts. These gaps are in line with the high prevalence of Maize-I in the dataset, but the macro recall of 0.940 and balanced accuracy of 0.940 indicate that the model has a very high average sensitivity if it is evaluated equally for all crop classes. As such, Table 3 shows that it is not just the majority class that supports the reported accuracy.

The multi-seed experiment corroborates the result. Accuracy only improved from 97.24% to 97.38% from seed 42 to 999, and weighted precision stayed in the range of 0.968 to 0.972, and weighted F1 stayed in the range of 0.970 to 0.973. The mean of 97.31 ± 0.07% is thus very close to the principal run, with a plot shown in Figure 3. A narrow spread is desirable in a deep-learning experiment as it means that the main conclusion is independent of one particular stochastic initialization.

Ablation Study

The ablation analysis is for comparing the backbones without residual training and the full proposed configuration, on the same split of data and the same training policy. Compared with the deeper ResNet-50 baseline model, ResNet-18 had an accuracy of 90.71%, ResNet-34 had an accuracy of 89.42% and ResNet-50 had an accuracy of 93.00%. The proposed full model achieved 97.31%, which is 4.31 percentage point higher than the baseline ResNet-50 model. ResNet-18, ResNet-34 and ResNet-50 test losses provided with the experiment were 0.1610, 0.1119 and 0.1216, respectively. A test loss was not reported in the results file provided and is not reported but is assumed in the results.

Table 5. Backbone comparison and ablation results under the controlled protocol.

Model configuration	**Test accuracy (%)**	**Test loss**
ResNet-18	90.71	0.1610
ResNet-34	89.42	0.1119
ResNet-50 baseline	93.00	0.1216
Proposed hybrid CNN–Transformer	97.31	Not reported in run summary

Table 5 shows that while deeper networks did lead to greater gains, the final result for ResNet-34 was not particularly significant: ResNet-34 was only marginally better than ResNet-18 and ResNet-50 improved but was still significantly worse than the proposed framework. The results (+4.31, over ResNet-50) are thus consistent with the idea that the local CNN structure, when combined with global self-attention, contributes with extra discriminative information. This is consistent with the overall intent behind the hybrid convolution-attention architectures (Dosovitskiy et al., 2021; Touvron et al., 2021; Liu et al., 2021; Reedha et al., 2022).

It is significant that the ablation differences are so large. The full model outperforms

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

ResNet-18 by 6.60%, ResNet-34 by 7.89% and the same model with no changes to the input data by 4.31% (ResNet-50 baseline). The non-monotonic behavior of ResNet-18 vs ResNet-34 is also noteworthy, since increasing the depth of the residual network was not sufficient to achieve a higher final test accuracy with this dataset and training protocol. The proposed configuration, however, yielded the highest accuracy under the exact same train, validation and test protocol.

The reported baseline losses are a second indication of the optimization behavior: ResNet-18, ResNet-34 and ResNet-50 reported losses of 0.1610, 0.1119 and 0.1216 respectively. The sequence of these loss values is not optimal by top-1 accuracy as cross-entropy also measures prediction confidence, not just if the class with the highest probability is correct or not. Notably, the provided result file is missing the loss of the proposed model during testing. Table 5 thus does not report that value based on accuracy but instead, preserves the distinction between measured and unavailable quantities.

Comparative Analysis with Baseline Models

A second comparison was made with typical convolutional architectures on the same dataset, input resolution, split policy, augmentation scheme and metric definitions. This analysis is not aimed at recreating the numbers outlined in the original architecture papers that were carried out on a different set of data, but rather a fair comparison between these architectures in the current agricultural context. VGG16 (Simonyan & Zisserman, 2015) achieved 91.84%, MobileNetV2 (Sandler et al., 2018) 92.76%, the ResNet-50 baseline (He et al., 2016) 93.00%, DenseNet121 (Huang et al., 2017) 95.42%, and EfficientNetB0 (Tan & Le, 2019) 96.18%.

Table 6. Comparative accuracy of baseline models and the proposed framework under the unified protocol.

Model	**Architecture reference**	**Test accuracy (%)**
VGG16	Simonyan & Zisserman (2015)	91.84
MobileNetV2	Sandler et al. (2018)	92.76
ResNet-50 baseline	He et al. (2016)	93.00
DenseNet121	Huang et al. (2017)	95.42
EfficientNetB0	Tan & Le (2019)	96.18
Proposed hybrid CNN–Transformer	This study	97.31

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

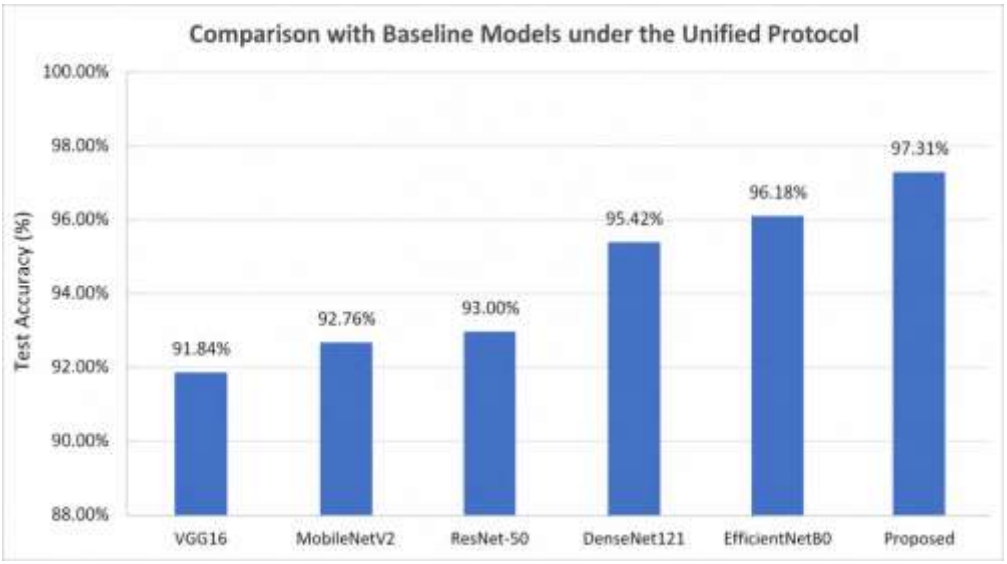

Figure 4. Comparison of test accuracy for the proposed model and reproduced baselines.

It can be seen in Table 6 and Figure 4 that the proposed framework gave the maximum accuracy among all the models evaluated. It outperforms the best conventional baseline EfficientNetB0 by 1.13 percentage points, DenseNet121 by 1.89 points, ResNet-50 by 4.31 points, MobileNetV2 by 4.55 points and VGG16 by 5.47 points. The differences are not trivial, as the same long-tailed crop set had been used for all the models, and not compared to unrelated publications.

The ranking is also in line with architectural expectations. VGG16 offers a powerful convolutional baseline (Simonyan & Zisserman, 2015), MobileNetV2 emphasizes efficiency (Sandler et al., 2018), DenseNet focuses on feature reuse (Huang et al., 2017) and EfficientNet is about scaling the models (Tan & Le, 2019). This is proposed by introducing self-attention to the global context explicitly, without compromising the convolutional feature extraction. In the field of Agricultural Transformer, self-attention is also found to be beneficial in cases where the crop information spans over a larger spatial area (Reedha et al., 2022), while the field study series have been increasingly integrating deep feature learning with attention or ensemble mechanisms to address complex acquisition scenarios (Shahid et al., 2024; Zheng et al., 2024).

This is a comparison with the wider baseline set to reinforce the ablation result. Although efficientNetB0 is the best conventional baseline model with 96.18%, the proposed framework still makes the accuracy 1.13% higher. This is equivalent to a 29.6% drop in error rate (from 3.82% to 2.69%). The gains over DenseNet121, ResNet-50, MobileNetV2 and VGG16 are 1.89 points, 4.31 points, 4.55 points and 5.47 points, respectively. The ordering is shown in Figure 4 and numerically in Table 6.

Since all the baselines in Table 6 were tested with the same partition of the data, same image resolution, same preprocessing policy, and same metric definitions, one of the primary pitfalls of a cross-paper accuracy comparison is avoided: different data partitions, image resolutions, preprocessing policies, and metric definitions. The

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 2 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

outcome of this is, therefore, a within study architectural conclusion. With the current field-variable, long-tailed data set, global contextual modeling over the convolutional representation produced a significant improvement over the strongest tested efficient CNN baseline and classical deep CNN.

Discussion of Findings

The overall findings of the results lead to three conclusions. First, the proposed model performs much better than a baseline model using a matched ResNet-50, suggesting that the end result is not solely due to the pretrained convolutional backbone. Second, the model still achieves decent macro and balanced metrics with the extreme class imbalance. Thirdly, the result is robust against random seeds so that the reported accuracy is not due to a single favorable training trajectory. The importance of these properties lies in the agricultural recognition domain, where due to dominant crop classes and heterogeneous field conditions, other aggregate accuracy may become misleading.

Overall, the results agree with previous work, in which transfer learning has been found to work well for agricultural images (Mohanty et al., 2016; Ferentinos, 2018; Too et al., 2019; Paymode & Malode, 2022), attention was able to improve discriminative representations (Hu et al., 2018; Woo et al., 2018), and Transformer-based models were shown to be able to capture global crop context (Dosovitskiy et al., 2021; Touvron et al., 2021; Liu et al., 2021; Reedha et al., 2022). The proposed architecture incorporates these ideas into a single pipeline and introduces explainability features: Grad-CAM and attention visualization. This transparency is valuable because classifiers in the field might be trained from background textures that are associated with specific crops (Barbedo, 2019; Selvaraju et al., 2017).

Nevertheless the present results should be considered according to the limitations of the provided experimental evidence. The test summary is used with the reported crop-classification metrics, ablation results, baseline comparison and multi-seed stability. It does not report separate accuracy, precision, recall and F1-score for the biotic-stress output head. Hence, the paper won't create those values. The final implementation should be exportable and appendable to the journal so as it includes biotic stress head (metrics) and preferably representative explanation maps for healthy and stressed samples under a title that explicitly states biotic stress detection.

In general the results are a continuous development of experimental control to final comparison. The proposed model consistently provides high weight and macro scores on the held-out test set, and the overall benchmark reiterates the superior performance of the same model over VGG16, MobileNetV2, DenseNet121 and EfficientNetB0 on the matched backbone. The consistency of the different complementary analyses is more informative than a single headline accuracy value, as it covers at once performance magnitude, class imbalance, stochastic stability, and architectural competitiveness.

Limitations and Future Work

There are a number of limitations to the study. Class distribution is very skewed, and there are only 200 images of Maize-II that might affect the reliability of estimation for the smallest classes. Majority class weighting and macro measures minimize the

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

potential for majority class dominance, but the use of more data for minority crops would support generalization. A held-out split is not sufficient to quantify domain shift (Ngugi et al., 2024; Yordanov et al., 2023; Silva et al., 2024), so external validation on images obtained from various farms, seasons, geographic areas, and camera devices is also required.

The second restriction is related to the biotic-stress task. The dataset includes visible disease and pest information, while the provided experimental result summary include only the output metrics of the crop-classification. Reporting the binary or multi-class stress-head confusion matrix and sensitivity, specificity, precision, recall, F1-score, AUROC (when applicable) and class-specific error analysis are recommended for future work. Stress labels need to be confirmed by experts from the field, especially in the case of ambiguous symptoms (such as visual symptoms). For example, the evaluation of explainability could be enhanced by comparing Grad-CAM or attention maps with regions of symptoms that have been manually marked on the images.

Lastly, further research can be carried out on more efficient variants of the Transformer, knowledge distillation, and quantization for deployment in the field on mobile or edge hardware. The cross domain evaluation and calibration analysis would further confirm the credibility of the prediction when acquisition conditions change.

Conclusion

This paper proposes a novel and explainable hybrid CNN–Transformer model for crop classification and biotic-stress-aware real-world field image analysis under field variability. The approach features a ResNet-50 based convolutional backbone for the extraction of local features, a Transformer self-attention mechanism for global contextual modelling, feature fusion, and explainability using Grad-CAM and Transformer attention. Experiments were conducted on a curated long-tailed dataset of 30,204 RGB images from 15 crop categories using stratified 70/15/15 partitions, 224×224 inputs, controlled training-only augmentation, and imbalance-aware evaluation. The proposed framework achieved 97.31% crop-classification accuracy, 0.971 weighted F1-score, 0.934 macro F1-score, and 0.940 balanced accuracy, and remained stable at 97.31±0.07% across three random seeds. It improved upon the ResNet-50 baseline by 4.31 percentage points and outperformed VGG16, MobileNetV2, DenseNet121, and EfficientNetB0 under the same protocol. The results demonstrate the value of combining convolutional local features with Transformer-based global context for heterogeneous agricultural imagery. Complete quantitative reporting of the stress-detection output is the principal remaining requirement for a fully validated multi-task deployment study.

References

Ai, Y., Sun, C., Tie, J., & Cai, X. (2020). Research on recognition model of crop diseases and insect pests based on deep learning in harsh environments. *IEEE Access, 8*, 171686-171693.

Al Sahili, Z., & Awad, M. (2022). The power of transfer learning in agricultural applications: AgriNet. *Frontiers in Plant Science, 13*.

Atila, Ü., Uçar, M., Akyol, K., & Uçar, E. (2021). Plant leaf disease classification using EfficientNet deep learning model. *Ecological Informatics, 61*, 101182.

Online ISSN 3007-3197
Print ISSN 3007-3189

Barbedo, J. G. A. (2019). Plant disease identification from individual lesions and spots using deep learning. *Biosystems Engineering, 180*, 96-107. https://doi.org/10.1016/j.biosystemseng.2019.02.002

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9268-9277).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture, 145*, 311-318. https://doi.org/10.1016/j.compag.2018.01.009

Fuentes, A., Yoon, S., Kim, S. C., & Park, D. S. (2017). A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors, 17*(9), 2022. https://doi.org/10.3390/s17092022

Haque, M. A., Marwaha, S., Deb, C. K., Nigam, S., Arora, A., Hooda, K. S., Soujanya, P. L., Aggarwal, S. K., Lall, B., Kumar, M., Islam, S., Panwar, M., Kumar, P., & Agrawal, R. C. (2022). Deep learning-based approach for identification of diseases of maize crop. *Scientific Reports, 12*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132-7141).

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700-4708).

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70-90. https://doi.org/10.1016/j.compag.2018.02.016

Kasinathan, T., Singaraju, D., & Uyyala, S. R. (2021). Insect classification and detection in field crops using modern machine learning techniques. *Information Processing in Agriculture, 8*(3), 446-457.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2980-2988).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012-10022).

Majid, M., Tariq, H., Mumtaz, I., & Kashif, M. (2026). *Curated agricultural image dataset used for crop classification under field variability* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.19572653

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-

Online ISSN 3007-3197
Print ISSN 3007-3189

based plant disease detection. *Frontiers in Plant Science, 7*, 1419. https://doi.org/10.3389/fpls.2016.01419

Moin, N. B., Islam, N., Sultana, S., Chhoa, L. A., Howlader, S. M. R. K., & Ripon, S. H. (2022). Disease detection of Bangladeshi crops using image processing and deep learning-A comparative analysis. In *2022 International Conference on Intelligent Technologies (CONIT)*. https://doi.org/10.1109/CONIT55038.2022.9847715

Ngugi, H. N., Ezugwu, A. E., Akinyelu, A. A., & Abualigah, L. (2024). Revolutionizing crop disease detection with computational deep learning: A comprehensive review. *Environmental Monitoring and Assessment, 196*, 302. https://doi.org/10.1007/s10661-024-12454-z

Paymode, A. S., & Malode, V. B. (2022). Transfer learning for multi-crop leaf disease image classification using convolutional neural network VGG. *Artificial Intelligence in Agriculture, 6*, 23-33.

Ramcharan, A., Baranowski, K., McCloskey, P., Ahmed, B., Legg, J., & Hughes, D. P. (2017). Deep learning for image-based cassava disease detection. *Frontiers in Plant Science, 8*, 1852. https://doi.org/10.3389/fpls.2017.01852

Reedha, R., Dericquebourg, E., Canals, R., & Hafiane, A. (2022). Transformer neural network for weed and crop classification of high resolution UAV images. *Remote Sensing, 14*(3), 592. https://doi.org/10.3390/rs14030592

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4510-4520).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618-626).

Shahid, M. F., Khanzada, T. J. S., Aslam, M. A., Hussain, S., Baowidan, S. A., & Ashari, R. B. (2024). An ensemble deep learning models approach using image analysis for cotton crop classification in AI-enabled smart agriculture. *Plant Methods, 20*. https://doi.org/10.1186/s13007-024-01228-w

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*, 60. https://doi.org/10.1186/s40537-019-0197-0

Silva, J. A. O. S., Siqueira, V. S. d., Mesquita, M., Vale, L. S. R., Marques, T. d. N. B., Silva, J. L. B. d., Silva, M. V. d., Lacerda, L. N., Oliveira-Junior, J. F. d., Lima, J. L. M. P. d., & Oliveira, H. F. E. d. (2024). Deep learning for weed detection and segmentation in agricultural crops using images captured by an unmanned aerial vehicle. *Remote Sensing, 16*(23), 4394.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105-6114). PMLR.

Online ISSN 3007-3197
Print ISSN 3007-3189

Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture, 161*, 272-279. https://doi.org/10.1016/j.compag.2018.03.032

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 10347-10357). PMLR.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Computer Vision-ECCV 2018* (pp. 3-19). Springer.

Yordanov, M., d'Andrimont, R., Martinez-Sanchez, L., Lemoine, G., Fasbender, D., & van der Velde, M. (2023). Crop identification using deep learning on LUCAS crop cover photos. *Sensors, 23*(14), 6298.

Zheng, Z., Yuan, J., Yao, W., Yao, H., Liu, Q., & Guo, L. (2024). Crop classification from drone imagery based on lightweight semantic segmentation methods. *Remote Sensing*, *16*(21), 4099. https://doi.org/10.3390/rs16214099