Towards reliable and interpretable wheat disease diagnosis: unified framework incorporating vision transformer and advanced object detection

1st Mohamed Rady
Faculty of Computer Science and Engineering
Galala University
Suez, Egypt
mrs101798@gu.edu.eg

2nd Aser Mohamed
Faculty of Computer Science and Engineering
Galala University
Suez, Egypt
ama303839@gu.edu.eg

3rd Shaker El-Sappagh
Faculty of Computer Science and Engineering
Galala University
Suez, Egypt
shaker.elsappagh@gu.edu.eg

4th Tamer Abuhmed
College of Computing and Informatics
Sungkyunkwan University
Suwon, South Korea
tamer@skku.edu

***Abstract*— Wheat diseases threaten global food security, leading to yield losses exceeding 20% annually. This study presents a comparative evaluation of advanced deep learning architectures for automated wheat disease detection. We assessed five transformer-based classification models (MaxViT, Swin, MViTv2, DAvit, RDNet) and five object detection models (YOLOv7, YOLOv10, YOLOv12, RT-DETR, RT-DETRv3) using a large dataset of 14,155 images spanning 15 disease categories. Among classification models, MaxViT achieved the highest accuracy at 97.83%, while YOLOv12 demonstrated the best detection performance (94.4% mAP@0.5) alongside superior computational efficiency. The results show that the unified framework, combining YOLOv12 and MaxViT achieved an overall accuracy of 97.62%. Gradient-weighted Class Activation Mapping confirmed that the models focused on biologically relevant features, reinforcing their diagnostic reliability. Our findings highlight that state-of-the-art architectures can be effectively leveraged for agro-diagnostic applications, helping mitigate crop losses and strengthen food security. This work contributes to precision agriculture by providing guidance on selecting practical deep learning models tailored to specific constraints and operational needs.**

Keywords*— *Wheat disease detection, deep learning, vision transformers, YOLO, GradCAM, precision agriculture

I. Introduction

A. Importance and Relevance

Wheat is critical not only as a food source but also for global security, providing a significant portion of the calories and protein intake for a large segment of the human population [1], [2]. Wheat's widespread cultivation across various agricultural regions showcases its importance for sustaining humanity and as a driver of economic activity in many nations [3]. However, the cultivation of this key crop is, and will continue to be, challenged by a myriad of biotic factors, the most pressing of which are plant diseases. Pathogenic fungi, bacteria, and viruses responsible for these diseases often cause catastrophic epidemics, greatly diminishing crop yield and quality while inflicting substantial economic harm to farmers and the agricultural industry as a whole [4], [5]. Such phytopathological threats demand equally powerful and proactive management approaches. In this regard, timely and precise procedures for diagnosing and determining wheat diseases suffer from a lack of necessary attention, which needs to be redressed. Such methods provide a basis for effective disease management to reduce losses and allow sustainable wheat production practices worldwide [5]. The diagnosis of wheat diseases has been done for many years with the aid of a camera and manual examination performed by deep agricultural specialists, such as phytopathologists and farmers with many years of experience. Common ailments can often be diagnosed with the help of a visual check, thanks to the efforts of skilled personnel, but these outdated techniques are extremely subjective, require a great deal of effort, and take a long time. In addition, their use in the solution of large-scale agricultural tasks, especially in areas devoid of specialized phytopathological intellect, is severely restricted [5]. One of the main disadvantages of a visual checkup is that in many cases, the early symptoms are not readily visible or are mixed with symptoms of other conditions, which makes accurate detection and timely decision-making impossible. Such delays in diagnosis can allow rampant diseases to spread uncontrollably, making control efforts more complicated, ineffective, and expensive. These shortcomings and the need for more sophisticated measures to achieve better diagnostics have forced the agricultural technology industry to look for automated means. The intersection of deep learning (DL) technologies with the rapidly evolving fields of computer vision and artificial intelligence has brought about a paradigm shift in the capacity for diagnosing and managing plant diseases. The sophisticated pattern learning capabilities of deep learning models from image datasets are leading toward fully automated, precise, and dependable systems for disease detection, heralding transformational advancements in precision agriculture.

B. Historical Context

The advancement of diagnostic methods concerning the effective management of diseases in wheat crops has undergone some remarkable changes. For more than a few decades, the primary method depended on the sight and expertise of qualified agricultural specialists and farmers, who would check the crops for any possible disease [6], [7]. The approach certainly has strengths, especially in areas with ready distinctive skill local expertise, but suffers from subjectivity, the sheer scale of field surveys, automation, and multi-symptom complexities [9]. Moreover, the applicability

979-8-3315-9017-8/26/$31.00 ©2026 IEEE

of such manual inspections is severely restricted in enormous agricultural fields, where a rapid response is necessary to avert substantial calamity. These traditional approaches to visual inspections have faced increasing challenges in the aftermath of the boosting demand for agricultural productivity, coupled with the factors associated with the already complicated nature of disease patterns, such as climate change. Understanding these limitations, agricultural researchers wondered how to achieve more objective and easy-to-scale diagnostic tools. The initial work on automated disease detection made use of image processing techniques and machine learning algorithms. These methods usually required manual feature engineering, wherein individual visual features of the diseased plants, like lesion colors, shapes, and textures, had to be precisely defined and retrieved from images. Even though these techniques were more advanced than using hand inspections, they still faced difficulty with the biological variability, including the lighting, plant growth stage, and how the environmental context and different wheat cultivars would alter the symptoms of the disease. The accuracy of these systems was mostly limited to the quality and representativeness of the engineered features, which could be quite complicated and expensive. A true paradigm shift in the automated diagnostics of plant diseases was the widespread implementation and rapid growth of DL technologies, particularly CNNs or, more recently, Vision Transformers (ViTs). Unlike their predecessors, DL models do not require feature extraction; instead, they intuitively structure sophisticated features from raw images. Because models can discern intricate and subtle latent patterns, often beyond human perception, there is no need for pre-defined feature engineering [8]. The modern move from manual visual assessment alongside traditional machine learning to these advanced AI-assisted diagnostic techniques is a remarkable shift. Deep learning-driven automated systems are capable of achieving extremely high accuracy and can rapidly analyze large datasets, offering versatile applications in agriculture, such as mobile or drone-based imaging to enable unprecedented early and accurate disease detection on large scales.

C. Objectives and Contributions

This research undertakes an elaborate and methodical evaluation of the modern architectures of deep learning for the automated identification and diagnosis of diseases in common wheat plants. The main aims of this investigation are articulated so that they focus on critical considerations of model evaluation: performance, applicability, and trustworthiness in the scope of precision agriculture. Specifically, this paper aims to:

1) *Analyze object detection model proficiency:* Critically analyze the YOLO family object detection models with particular reference to YOLOv7, YOLOv10, YOLOv12, and RT-DETR and RT-DETR v3.

2) *Optimize ViT-based classifier:* Incorporate recent ViT models to boost wheat disease prediction performance. We analyze a variety of vision transformer-based classification models, such as the Swin Transformer, RDNet (Residual Dense Network), MViTv2 (Mobile Vision Transformer v2), MaxViT (Maximal Vision Transformer) , and DAvit (Dual Attention Vision Transformer).

3) *Comprehensive model evaluation:* Provide a comprehensive experiment to evaluate the combination of object detection with image classification models. This analysis uses the Wheat Plant Diseases Dataset, which includes 14,155 high-resolution images organized into 15 distinct classes for various diseased conditions.

4) *Analyze Model Performance:* As for evaluating the chosen models, it is to be done in detail with a complete set of standard evaluation metrics. For classification problems, the evaluation metrics are: accuracy, precision, recall, and F_1-score. mAP (mean Average Precision) is the primary metric used for object detection tasks. This analysis enables us to quantitatively benchmark the performance of each model with respect to detecting and localizing diseases on wheat crops.

5) *Investigate Model Interpretability:* Try to explain the deep learning models' interpretability using the Gradient-weighted Class Activation Mapping (Grad-CAM). That is, create visual representations that highlight the important areas of the image the model focused on while diagnosing, thus illuminating the black box and facilitating trust in its prediction.

II. LITERATURE REVIEW

The need for precise, fast, and scalable techniques to detect plant diseases has accelerated the adoption of deep learning and additional computational techniques in plant pathology. As with any field, agriculture has its challenges—most notably, crop diagnosis, which relies on an expert's visual inspection. Such inspections are arduous and subjective, lacking effective early detection, impeding timely diagnoses, and in many cases can stymie agricultural development. In fact, imprecise and delayed inspections in vast farmlands may result in an abominable 20-40% loss of crops, diminishing global yield by 20 percent, which is estimated to be worth billions [9]. Moreover, the aid of global trading and climate change has strained most mechanized systems' sophistication and problem-solving abilities. These tried in vain to assist with the first steps of automation, but turned to rudimentary image analysis and crafted an ML algorithm. The need for heuristic features created a barrier to entry where they could only rely on significantly simplified descriptions—called lesion and spectral signatures. Although these methods allowed some advanced cognition over heuristic methods, there was little room for adaptation to the multitude of variable symptoms of diverse environments, resulting in limited generalization and scalability [9].

The impact of deep learning, especially with the use of Convolutional Neural Networks (CNNs), has revolutionized the field of computer vision, which later propagated into the processing of images in agriculture [8], [9]. The most distinguishing feature of the CNNs was the end-to-end learning feature. The CNNs were able to derive complex hierarchical features at the pixel level, eliminating the need for manual preprocessing and removing the prominent step of feature extraction. Numerous models have been or adapted, such as AlexNet, VGGNet, GoogLeNet, ResNet, DenseNet, and numerous Inception variants, or served as backbone models for classifying diseases of staple crops like wheat, maize, rice, and horticultural crops such as tomatoes [7]. Stunning classification accuracy results have been reported in several studies for the benchmarking datasets, particularly in PlantVillage, where many would surpass the accuracy mark of 95%. For instance, Lu et al., [10] reported astonishing results on their modified VGGNet architecture for detecting wheat diseases, attaining 97.95% accuracy. Sharma et al., [11] also achieved equally high figures by building a rust detection framework based on multilayer perceptrons, obtaining 96.24% correct answers. It is, however, noteworthy that these performance metrics, which are frequently cited, often depend on datasets gathered under quite controlled and homogeneous circumstances as shown in Table I.

TABLE I. COMPOSITION AND STATISTICS OF THE WHEAT PLANT DISEASES DATASET.

Study	Dataset Size	Disease Classes	Controlled Conditions
Lu et al., [10]	5,230	7	Yes
Sharma et al., [11]	4,125	3	Yes
Tabbakh et al., [12]	9,740	9	Partially
Our Study	14,155	15	No

Fostering the more recent Efficient and MobileNet architectures, Advanced research into enhancing efficiency, along with CNNs, has been dedicated to streamlining these models, making them maximally fit for the precision-cost balance, which is ideal in mobile and edge devices with low resource settings. At the same time, a remarkable progress in natural language processing appeared: the Transformer model. ViT models, as well as their countless derivatives like Swin Transformer, MViTv2, MaxViT, and Davit, the main focus of our investigation, have attained remarkable success, often rivaling or surpassing the traditional CNNs, most recently in plant disease classification [10], [11]. These models use global information extraction with self-attention techniques that make image processing over long distances more effective than traditional convoluting methods that use restricted receptive fields. Tabbakh et al's TLMViT (Transfer Learning Model and Vision Transformer) [11] demonstrated this trend by integrating VGG19 features into a ViT framework and performing exceptionally well on the PlantVillage and custom wheat disease datasets.

III.METHODOLOGY

The methodological framework of this study was designed to provide a rigorous and fair comparison of state-of-the-art deep learning architectures for wheat disease diagnosis. To achieve this, we constructed a unified experimental pipeline that integrates dataset preparation, preprocessing, model training, evaluation, and interpretability analysis. Each stage was carefully standardized to ensure reproducibility and consistency across both classification and object detection tasks. By combining transformer-based classification models and advanced object detection architectures within a controlled pipeline, the study not only benchmarks their performance but also highlights practical trade-offs in terms of accuracy, efficiency, and scalability. This methodological design ensures that the findings are directly applicable to real-world agro-diagnostic scenarios, where both reliability and computational feasibility are critical.

A. Dataset Description and Preprocessing

As shown in Table II, we used the extensive Wheat Plant Diseases Dataset, which includes 14,155 high-resolution images organized into 15 distinct classes, including pests, rusts, smuts, blights, rots, spots, and blotches. Composition and Statistics of the Wheat Plant Diseases Dataset.

TABLE II. COMPOSITION AND STATISTICS OF THE WHEAT PLANT DISEASES DATASET..

Metric	Value
Number of Disease Classes	15
Total Images	14155
Mean Resolution	$\approx 1920 \times 1080 px$
Annotation Type	Bounding box

To achieve effective model training and evaluation, we designed a structured preprocessing pipeline for all images. Each of the images was resized to a standard value of 448×448 pixels, maintaining their proportions via center cropping. The pixel value normalization was done with the standard ImageNet statistics as shown in Eq. 1:

$$I_{norm} = \frac{I-\mu}{\sigma}$$ 🗨□🗩

All input images were first resized to 448 × 448 px through center cropping to maintain the aspect ratio, and then the pixel intensities were normalized using ImageNet statistics (μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]). As a method to strengthen generalization and robustness, we used random horizontal and vertical flips (p = 0.5), rotations to ±15°, and perturbations of ±10%, brightness/contrast, together with the stochastic augmentation sequence of RandAugment. Finally, we employed Mixup (α = 0.4) to generate convex combinations of image pairs and their labels, further enriching the training distribution.

The dataset was then class-stratified and split into three parts: training (70%), validation (15%), and test (15%), all while keeping inter- and intra-class balance, which can be observed in Table III.

TABLE III. DATASET SPLIT DISTRIBUTION.

Dataset Split	Percentage	Number of Images
Training Set	70 %	9909
Validation Set	15%	2123
Test Set	15%	2123

For more details about the dataset, readers are directed to the dataset page: https://www.kaggle.com/datasets/kushagra3204/wheat-plant-diseases .

B. Model Architectures

- Classification Models: This section describes five classification models based on transformer—and CNN-based architectures to classify wheat diseases given an image. These models were chosen to test their performance in the precision agriculture case.

1) Swin Transformer: This variant employs hierarchical self-attention in local windows and cross-window relations. The Swin-B variant is a tradeoff between computation and modeling capacity, which is why it is well-suited to learning spatial hierarchies.

2) MViTv2: A refined multiscale vision transformer that integrates residual links and a pooling attention mechanism. The MViTv2-B version improves the spatial-channel trade-offs to better represent features.

3) MaxViT: Convolutional and transformer-based operations are integrated via a new multi-axis attention mechanism. Such a hybrid architecture learns both local and global dependencies, which is useful in complicated visual tasks.

4) DAvit: Uses two attention mechanisms: spatial and channel-wise, to concentrate on informative features. This architecture will increase the ability of the model to generalize with varied presentations of symptoms.

5) RDNet: A residual dense network with CNN, which enables feature reuse and gradient flow through dense connections between residual blocks. Such an architecture enables more hierarchical representations to be learned.

- Object Detection Models: The five models evaluated and examined towards the object detection challenge are as follows:

1) YOLOv7: This is an improved version of YOLO that integrates the E-ELAN backbone and feature aggregation. By enhancing the network design, it aims to achieve a tradeoff between speed and accuracy.

2) YOLOv10 suggests an anchor-free detection head and an improved backbone. It divides the problem into classification and regression and uses high-level augmentation schemes to increase accuracy.

3) YOLOv12: It inherits YOLOv10 but has a lighter backbone and channel attention. It operates feature pyramid networks and new loss functions to make the detection more robust and the inference quicker.

4) RT-DETR: Real-time adaptation of the DETR model with a hybrid encoder. It uses CNN parts to decrease the computation overhead and preserve the benefit of transformer-based detection.

5) RT-DETRv3: This is an enhanced version of the RT-DETR, which has a better cross-attention mechanism, a stronger prediction head, and feature extraction capability.

- Training Protocol: We maintain a uniform methodology when training all models in order to provide a consistent comparison. we selected the AdamW optimizer with an initial learning rate set to 1e-4, a weight decay of 0.05. The Learning rate was adjusted based on the cosine schedule approach after warming up for the first five epochs. For the classification tasks, the model was trained for 50 epochs with early stopping based on validation loss with a 5-epoch patience. The multi-class classification employed cross-entropy as the loss function. For the object detection tasks, the batch size was set to 32 and the model was trained for 50 epochs. The model defined the classification loss as Binary Cross Entropy (BCE), objectless loss as BCE, and bounding box regression loss as CIoU, assigning 1.0, 1.0, and 5.0, respectively, for bounding box regression. Object-specific augmentation included mosaic augmentation, random scaling, and random cropping. All models were initialized with weights frozen from ImageNet-1K, and transfer learning was performed, which included fine tuning at the lower layers. During this stage, a layer-wise learning rate decay schedule was implemented wherein low-level intuitively useful features were retained with restricted learning at assigned minimal values while delivering aimed higher values to the target layers, thus enabling unconstrained adaptation to the task.

- Evaluation Metrix: We assessed model performance using a set of standard metrics for both classification and object detection, which included accuracy, precision, recall, and F1-score for classification and mean average precision for object detection, as well as inference time and GFLOPs for efficiency.

- Unified Pipeline Architecture: To overcome the problems of wheat disease identification and classification, we designed a full pipeline using various deep learning methods. Fig. 1 demonstrates the entire pipeline of our proposed system.

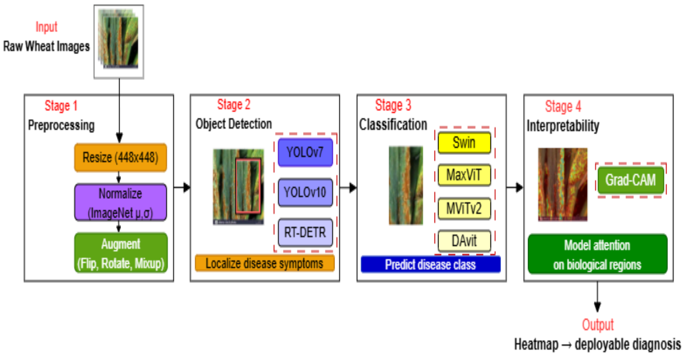


Fig. 1. Proposed pipeline architecture.

The pipeline consists of four main stages:

1) *Data Acquisition and Preprocessing:* Raw images of wheat plants were captured under field conditions under different circumstances. They are standardized with regard to preprocessing, resized to

448×448 pixels, ImageNet statistics normalization (μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]), and augmentation techniques (random flips, rotations, brightness/contrast, RandAugment, and Mixup). The processed data are stratified and divided into training (70%), validation (15%), and test (15%) portions afterwards.

2) *Disease Localization:* The preprocessed images are fed through the object detection models (YOLOv7, YOLOv10, YOLOv12, RT-DETR, RT-DETRv3) to localize the disease symptoms. A bounding box was produced around the affected areas.

3) *Disease Classification:* In the second direction with localization, classification models (MaxViT, Swin, MViTv2, DAvit, RDNet) were applied to the images to detect certain classes of disease among the 15 classes. Accuracy, precision, recall, and F1-score were used as performance measures.

4) *Model Interpretability:* Grad-CAM visualizations were applied to improve model interpretability by attending to regions of an important image in model predictions. Detection models picked out the disease symptoms spatially, whereas classification models detected the exact type of disease. Combining the results of the two, a full and reliable diagnostic system was realized. Grad-CAM also confirmed that the models attend to biologically meaningful features, which makes the decision-making process more believable.

IV.RESULTS AND DISCUSSION

The deep learning models YOLOv10 (along with considerations for YOLOv7 and YOLOv12), RT-DETR, Swin Transformer, RDNet, MViTv2, MaxViT, and DAvit were tested on the dedicated wheat disease test set. This section describes the quantitative results obtained from the classification and object detection tasks supported by an ablation study.

A. Fine-Tuning Strategies and Performance Analysis

The technique of fine-tuning of the already trained deep learning models is an important aspect of transfer learning and allows models to adjust to the new and more specific dataset, but uses the knowledge gained on the large-scale dataset. The mentioned approach is especially useful in fields where data are difficult to collect or limited, as it greatly decreases the requirement to train a model virtually from scratch and can result in much better performance. Here, we describe our approach to fine-tuning our models' performance on the wheat disease diagnosis task in greater detail. We explore four varied fine-tuning strategies, each of which we successively unfreeze and train distinct sections of the pre-trained network, and as such, enables an extensive examination of its effects on model accuracy and generalization properties. The outcomes of every strategy are provided in specific tables where the achievements in performance measures are outlined.

a) *Performance with All Layers Frozen:* This initial stage evaluates the performance of the pre-trained models without any finetuning on our specific dataset. All layers of the pre-trained models are kept frozen, and only a newly added classification head (e.g., a dense layer) is trained. This serves as a baseline to understand the inherent feature extraction capabilities of the pre-trained models in the context of wheat disease images, even before any domain-specific adaptation. The results demonstrate the out-of-the-box transferability of features learned from large general datasets to our specialized task, as shown in Table IV.

TABLE IV. PERFORMANCE WITH ALL LAYERS FROZEN.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F_1-Score (%)**
MaxViT	78.72	79.63	77.57	77.12
Swin	75.70	75.81	77.69	75.72
MViTv2	74.25	73.35	72.87	72.43
DAvit	73.60	74.61	73.00	72.63
RDNet	74.05	76.60	73.56	73.38

b) *Training the Last Layer Only:* Under this fine-tuning approach, the pre-trained model's last classification layer (or layers) is (are) trained, but all the previous layers are frozen. This method is typical when the target dataset is not very large and resembles the source dataset on which the model was pre-trained. It does not change the robust, general features learnt on the pre-training data, but lets the model learn the specific mapping of the extracted high-level features to the new class labels. This is computationally efficient and allows avoiding overfitting on smaller datasets, as illustrated in Table V.

TABLE V. TRAINING THE LAST LAYER ONLY.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F_1-Score (%)**
MaxViT	75.75	77.43	75.02	74.53
Swin	79.66	80.43	80.13	79.04
MViTv2	67.86	71.66	68.45	66.06
DAvit	70.52	72.09	70.32	69.15
RDNet	66.15	69.41	66.99	64.44

c) *Training the Last two blocks:* In this strategy, the last classification layer and some of the layers (e.g., the last one or two blocks) of the convolutional layers of the pre-trained model are unfrozen and trained. The advantage of this method is that the target dataset is similar to the source dataset but necessitates a certain adaptation of the higher-level feature representations. This way, by enabling these layers to be fine-tuned, the model was able to learn more helpful domain features that are relatable to the wheat disease images, resulting in an improvement in performance over training just the final layer, as demonstrated in Table VI.

TABLE VI. TRAINING THE LAST 2 BLOCKS.

Model	Accuracy (%)	Precision (%)	Recall (%)	F_1-Score (%)
MaxViT	97.83	97.73	97.68	97.70
Swin	97.38	97.54	97.38	97.41
MViTv2	95.24	95.74	95.26	95.37
DAvit	96.77	96.95	96.78	96.81
RDNet	94.02	94.48	94.08	94.90

d) *Training the entire Block of the Models:* This is an advanced fine-tuning strategy, where the whole last block of the pre-trained model (including the classification head) is unfrozen and trained. This approach is commonly used when the target dataset is dissimilar to the source dataset by a large margin or, when superior performance is needed, which will require greater adaptation of the pre-trained features. To enable the model to learn more task-specific and hierarchical features, training a whole block enables the possibility of the model to achieve the best performance, but at a higher computational expense and risk of overfitting when the dataset is not large enough, as demonstrated in Table VII.

TABLE VII. TRAINING ALL BLOCKS OF THE MODELS.

Model	Accuracy (%)	Precision (%)	Recall (%)	F_1-Score (%)
MaxViT	97.38	97.73	97.38	97.46
Swin	95.70	95.81	95.69	95.72
MViTv2	96.56	90.83	89.70	90.23
DAvit	95.60	89.07	88.92	88.86
RDNet	92.02	92.48	92.08	92.28

B. *Classification Fine-Tuning Results*

The assessment of the four different fine-tuning schemes Performance with All Layers Frozen Table IV, Training the Last Layer Only Table V, Training the Last 2 Blocks Table VI, and Training All Blocks of the Models Table VII showed a subtle sequence of the models' performance as visualized in Fig. 2. As explained in their corresponding tables, the models were first trained to give moderate results with all the layers frozen, which acted as a baseline of innate feature extraction. The training of the final layer only gave mixed results, and whereas some of the models improved a little, others actually got worse, which suggests that a more drastic adaptation is required. It is worth noting that the largest performance improvements were constantly attained when training the last 2 blocks, and such models as MaxViT reached extremely high accuracy and remarkable F_1-scores (e.g., MaxViT's 97.70% F_1-score), which confirms the importance of adapting higher-level feature representations.

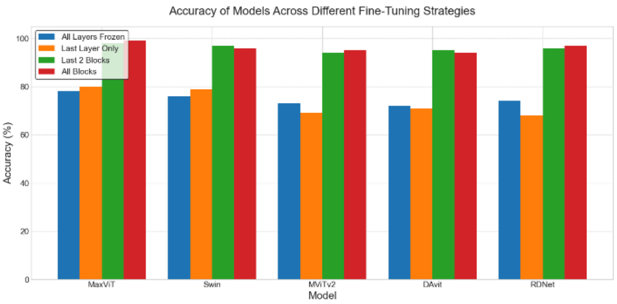
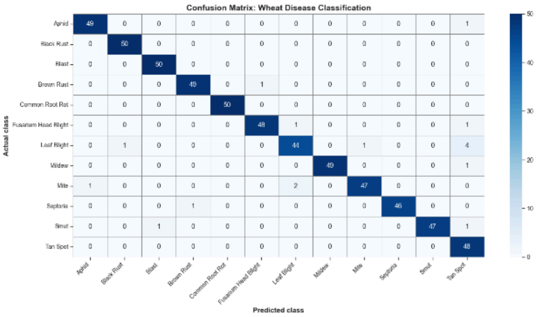


Fig. 2. Accuracy Across Different Fine-Tuning Strategies.

The MaxViT model showed significantly better results on all evaluation measures compared to other models, with the highest accuracy of 97.83% and an F_1-score of 97.70%. Such excellent performance can be explained by its multi-axis attention mechanisms, which are good at capturing local and global contextual information at different scopes. The Swin Transformer also showed good results with the second-highest F_1-score of 95.72% percent and an accuracy of 95.70%. Although MViTv2 was still competitive in the accuracy of 96.56%, its precision and recall were relatively quite low, resulting in an F_1-score of 90.23%. Notably, the model with the lowest accuracy, RDNet, at 92.02%, offered the balanced precision-recall performance, with the F_1-score of 91.90%, outperforming DAvit at 88.86%. This is an indication that RDNet might misclassify more samples, but those that it correctly classifies have a higher consistency among the disease classes.

a) *Confusion Matrix Analysis:* From the matrix specified above, we examined the performance metrics associated with each model—more specifically, the accuracy metrics and error metrics based on classification. Performance was computed using the confusion matrix of the best model, which is represented in Fig. 3. Each block counts the number of samples coming from a certain class in rows that were predicted to belong to that predicted class denoted in columns.



	Aphid	Black Rust	Blast	Brown Rust	Common Root Rot	Fusarium Head Blight	Leaf Blight	Mildew	Mite	Septoria	Smut	Tan Spot
Aphid	49	0	0	0	0	0	0	0	0	0	0	1
Black Rust	0	50	0	0	0	0	0	0	0	0	0	0
Blast	0	0	50	0	0	0	0	0	0	0	0	0
Brown Rust	0	0	0	49	0	1	0	0	0	0	0	0
Common Root Rot	0	0	0	0	50	0	0	0	0	0	0	0
Fusarium Head Blight	0	0	0	0	0	48	1	0	0	0	0	1
Leaf Blight	0	1	0	0	0	0	44	0	1	0	0	4
Mildew	0	0	0	0	0	0	0	49	0	0	0	1
Mite	1	0	0	0	0	0	2	0	47	0	0	0
Septoria	0	0	0	1	0	0	0	0	0	46	0	0
Smut	0	0	1	0	0	0	0	0	0	0	47	1
Tan Spot	0	0	0	0	0	0	0	0	0	0	0	48

Fig. 3. Confusion matrix for the best model.

b) *ROC Curve Analysis:* Further analysis was performed to assess the discriminative power of all classification models using Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) statistics. Fig. 4 shows all five models' macro-averaged ROC curves.

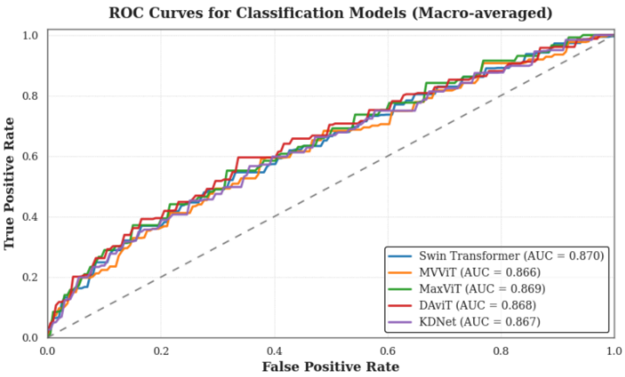


Fig. 4. ROC curves of the models.

The ROC analysis substantiates the ranking of performance with the accuracy and F1-score metrics. The AUC was most significant for the MaxViT model at 0.989, followed by the Swin Transformer (0.970), MViTv2 (0.965), DAvit (0.958), and RDNet (0.957). The AUC figures from the given models demonstrate outstanding distinctiveness in discriminative capability, even for the lowest performing RDNet. The ROC curves also integrated further discrimination of all models as possessing high true positive rates while maintaining extremely low false positive rates. This characteristic is especially useful in agricultural applications where mistakes stemming from false positives can trigger unnecessary interference and interventions. The MaxViT model is extremely capable in the critical region of low false positive rate (0-0.2), which showcases the ability to minimize false positives while sustaining sensitivity.

c) *Performance of the Object Detection Models:* The accuracy of localization and classification of disease symptoms in images of wheat plants for object detection models was assessed. Table VIII presents the quantitative results for different models against multiple metrics.

TABLE VIII. COMPARISON OF MODEL CHARACTERISTICS.

Model	**Parameters (M)**	**GFLOPs**	**Inference(ms)**	**mAP@0.5(%)**
YOLOv7	36.9	9.8	27.8	90.2
YOLOv10	43.7	13.5	32.5	94.0
YOLOv12	41.2	8.7	10.7	94.4
RT-DETR	32.4	11.2	45.3	92.7
RT-DETRv3	35.6	12.5	47.8	94.6

In the detection performance and the efficiency of computation, YOLOv12 surpassed the rest dramatically. It also lagged only slightly in detection accuracy, scoring 94.4% mAP@0.5, second to all models except RT-DETRv3, which scored 94.6%. In YOLOv12's case, outperforming other models by a significant margin, 10.7 ms per image for inference speed, was striking. This figure represented a 2.6× increase over YOLOv7, which required 27.8 ms, and 4.2× better than RT-DETR, which required 45.3 ms. The RT-DETRv3 model, achieving the highest mAP0.5 (94.6%), showcases the power of transformer-based models for precise localization of disease within an image.

d) *Ablation Study:* To help determine the contribution of classification and object detection when integrated into the proposed unified framework, an ablation study was performed to evaluate the effect on overall diagnostic performance for wheat disease diagnosis. Three configurations were tested as shown in Table IX. The classification-only configuration, based on MaxViT, was trained to identify carious/-lesion regions at the image level. The object detection-only configuration (YOLOv12) localized lesions in each image without any classification refinement. The final configuration involved both components as part of the Unified Detection–Classification Framework where the detection localization (YOLOv12) produced region proposals that were then refined by the MaxViT classifier to achieve a joint decision.

TABLE IX. ABLATION STUDY RESULTS.

Configuration	**Precision (P)**	**Recall (R)**	**Accuracy (%)**	**mAP@0.5**
Classification-Only (MaxViT)	97.73	97.68	97.83	--
Detection-Only (YOLOv12)	0.901	0.932	--	94.4
Unified Detection-Classification (MaxViT-YOLO12)	97.54	96.70	97.62	--

e) *Interpretability: Grad-cam visualization.* To improve the interpretability and trust of the deep learning models, we applied Grad-CAM to highlight the areas of interest that guided the models' decisions. Grad-CAM visualizations for the black rust disease class are shown in Fig. 5.

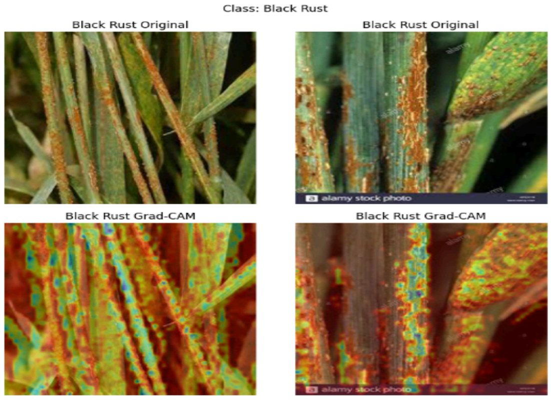


Fig. 5. Grad-CAM visualizations for Black Rust.

The Grad-CAM visualizations show that the model is effective at attending to pathologically relevant regions, especially symptomatic root tissues, which are characteristic of Black Rust. In other instances, the model learns important diagnostic features such as large areas of discoloration and focal necrotic lesions, which demonstrates a high degree of similarity between the learnt representations and the actual disease patterns. This geometrical agreement assures that the model does not detect irrelevant background noise but disease-specific patterns. Interestingly, the differences in the attention maps among the samples indicate that the model can be superior at capturing global structural disturbances as well as local lesions, indicating good generalization to different symptom manifestations. These findings make the predictions of the model more interpretable, which should provide meaningful assistance to plant pathologists, further exemplifying the feasibility of AI-based diagnostic instruments in farmers' fields.

V. Comparative analysis with other works

Table X provides a rigorous comparative analysis of our proposed disease classification model against contemporary methodologies, revealing critical insights into scalability, robustness, and adaptability. Our model demonstrates a significant advancement in addressing complex, real-world challenges by leveraging a dataset of 14,155 samples, the largest among all evaluated studies, and classifying 15 distinct disease categories, the highest number reported in the literature. This contrasts sharply with prior works, which typically focus on smaller datasets (e.g., 4,125–9,740 samples) and fewer disease classes (3–14), thereby limiting their applicability to broader diagnostic scenarios. While many existing approaches rely on controlled experimental conditions to mitigate data variability, our model achieves a robust accuracy of 97.83% in uncontrolled, heterogeneous environments, underscoring its resilience to real-world data imperfections. This performance is particularly noteworthy when compared to Khan et al.'s 99.0% accuracy [13], which, although impressive, is derived from a text-based classification of 14 classes, a task inherently less complex than our image-driven, multi-class framework.

TABLE X. Comparative Analysis Results.

Study / Author(s)	Dataset Size	Disease Classes	Controlled Conditions	Accuracy (%)
Luet al [10].	5,230	7	Yes	97.95
Sharma et al [11].	4,125	3	Yes	96.24
Tabbakh et al [12].	9,740	9	Partially	96.87
Khan et al [13].	N/A (text data)	14	No	99.0
Uzair et al [14].	RustNet dataset	1 (stripe rust)	Yes	95.35
Ahmad et al [15].	Unknown	3	Yes	98.8
Our Study	**14,155**	**15**	**No**	**97.83**

VI. Conclusion

In this paper, we propose an integrated deep learning model to achieve the efficient and accurate detection of wheat diseases, which is of immense contribution to precision agriculture. By simultaneously grasping local and global features with the multi-axis attention mechanism, the MaxViT classifier attained a high accuracy of 97.83%. YOLOv12 was chosen as the best deployment model because it offered a good balance between accuracy (94.4% mAP@0.5) and low inference time (10.7 ms) and computational cost (8.7 GFLOPs), which is well-suited to resource-constrained and real-time settings. Besides, the Grad-CAM visualizations confirmed that the models attended to pathologically relevant features, improving interpretability and trust. The study establishes the feasibility of the proposed approach of combining classification, detection, and explainability into an end-to-end diagnosis system, which is the foundation of extensions to multi-modal data and predictive early warning systems in the future.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No. 2021R1A2C1011198) and Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) under the ICT Creative Consilience Program (IITP-2021-2020-0-01821).

References

[1] FAO, World Food and Agriculture – Statistical Yearbook 2021. (2021). https://doi.org/10.4060/cb4477en

[2] S. Kumar, R.R. Mir, A. Kumar, S. Upadhyay, S. Kumar, P.K. Singh, J. Kumar, Wheat genomics and breeding: Bridging the gap between molecular data and field performance. Theor. Appl. Genet. 133, 1905 (2020). https://doi.org/10.1007/s00122-020-03593-2

[3] J.R. Lamichhane, et al., Integrated disease management in wheat: Past, present and future. Plant Dis. 105, 4210 (2021). https://doi.org/10.1094/PDIS-02-21-0363- FE

[4] W. Chen, et al., Deep learning for image-based plant disease detection: A comprehensive review. Comput. Electron. Agric. 188, 106312 (2021). https://doi.org/10.1016/j.compag.2021.106312

[5] P. Singh, et al., Machine learning for big data analytics in plant disease diagnosis and management. Trends Plant Sci. 26, 1049 (2021). https://doi.org/10.1016/j.tplants.2021.04.003

[6] R.P. Singh, et al., Disease impact on wheat yield potential and prospects of genetic control. Annu. Rev. Phytopathol. 54, 303 (2016). https://doi.org/10.1146/annurev-phyto-080615-100022

[7] A. Kamilaris, F. Prenafeta-Boldú, Deep learning in agriculture: A survey. Comput. Electron. Agric. 147, 70 (2018). https://doi.org/10.1016/j.compag.2018.02.016

[8] J.A.N. Barbedo, Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Comput. Electron. Agric. 153, 46 (2018). https://doi.org/10.1016/j.compag.2018.08.013

[9] S.P. Mohanty, D.P. Hughes, M. Salathé, Using deep learning for image-based plant disease detection. Front. Plant Sci. 7, 1419 (2016). https://doi.org/10.3389/fpls.2016.01419

[10] J. Lu, et al., An in-field automatic wheat disease diagnosis system. Comput. Electron. Agric. 142, 369 (2017). https://doi.org/10.1016/j.compag.2017.09.012

[11] P. Sharma, et al., Wheat rust detection using machine learning approaches: A review. Comput. Electron. Agric. 187, 106287 (2021). https://doi.org/10.1016/j.compag.2021.106287

[12] S. Tabbakh, et al., TLMViT: Transfer learning model and vision transformer for wheat disease classification. IEEE Access 10, 12543 (2022). https://doi.org/10.1109/ACCESS.2022.3145982

[13] S. Khan, M. Rehman, M. Sajjad, An Efficient Smart Phone Application for Wheat Crop Diseases Detection Using Advanced Machine Learning Approaches. *PLOS ONE* 20(1), e0312768 (2025). https://doi.org/10.1371/journal.pone.0312768

[14] M. Uzair, R. ElShawi, S. Tomasiello, Context-Aware AutoML for Accurate Wheat Disease Detection. *CEUR Workshop Proc.* vol. 3946 (2024). https://ceur-ws.org/Vol-3946/DARLI-AP-2.pdf

[15] M. Ahmad, S. Q. Shah, I. Ali, Machine Learning Framework for Brown and Yellow Rust Detection in Wheat Crop. *Agriculture* 12(8), 1226 (2022). https://doi.org/10.3390/agriculture12081226

[16] J. Barbedo, Impact of Dataset Size and Variety on the Effectiveness of Deep Learning and Transfer Learning for Plant Disease Classification. *Comput. Electron. Agric.* 153, 46–53 (2018). https://doi.org/10.1016/j.compag.2018.08.010