Data-Efficient Wheat Disease Detection Using Shifted Window Transformer: Enhancing Accuracy, Sustainability, and Global Food Security

Muhammad Khubaib, Tanzila Kehkashan, Maha Abdelhaq, *Member, IEEE*, Muhammad Asghar Khan, *Senior Member, IEEE*, Muhammad Zaman, *Member, IEEE*, Imran Ashraf, Abdul Rehman, and Adnan Akhunzada, *Senior Member, IEEE*

***Abstract*—Wheat is an essential crop that plays a vital role in global food security, but is susceptible to a variety of diseases, which have the potential to drastically decrease crop productivity. Detection of disease at an early stage and in an accurate manner is crucial to minimize crop losses. This research presents a deep learning technique based on the Shifted Window (Swin) Transformer, a powerful attention-based model that effectively captures both local and global information for enhanced classification output. Unlike conventional CNN-based methods, which often face limitations in feature extraction, the Swin Transformer utilizes hierarchical attention mechanisms to improve disease detection accuracy. The proposed modelis trained on a dataset of 9,346 wheat leaf images, categorized into eight disease classes and one healthy class. Using Bayesian hyperparameter optimization, we tuned key parameters such as learning rates, batch sizes, and dropout rates, and achieved an accuracy of 99.3%. We also conducted comparative analyses with baseline self-attention, CNN-based feature extraction, and hybrid attention layers. To enhance interpretability, Grad-CAM visualization techniques were applied, confirming model reliability. This research advances precision agriculture by improving the efficiency of wheat disease identification and supporting sustainable farming through state-of-the-art deep learning methodologies.**

***Index Terms*—Wheat disease detection, deep learning, Swin-transformer, image classification, precision agriculture.**

Received 30 January 2025; revised 4 April 2025 and 29 April 2025; accepted 19 June 2025. Date of publication 23 June 2025; date of current version 7 November 2025. This work was supported by the Princess Nourah bint Abdulrahman University Researchers Supporting Project, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, under Grant PNURSP2025R97. *(Corresponding authors: Muhammad Zaman; Muhammad Asghar Khan; Adnan Akhunzada.)*

Muhammad Khubaib, Imran Ashraf, and Abdul Rehman are with the Faculty of Information Technology, The University of Lahore, Sargodha 40100, Pakistan (e-mail: imranashraf.yaseen@gmail.com; m.abdulrehman.shoukat@gmail.com).

Tanzila Kehkashan is with the Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia, and also with the Faculty of Information Technology, The University of Lahore, Sargodha 40100, Pakistan (e-mail: tanzila.kehkashan@gmail.com).

Maha Abdelhaq is with the Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia (e-mail: msabdelhaq@pnu.edu.sa).

Muhammad Asghar Khan is with the Department of Electrical Engineering, Prince Muhammad Bin Fahd University, Al Khobar 31952, Saudi Arabia (e-mail: m.asghar@ieee.org).

Muhammad Zaman is with the Department of Computer Science, COMSATS University Islamabad, Islamabad 45600, Pakistan, and also with the Department of Computer Science and IT, Superior University, Sargodha 40100, Pakistan (e-mail: muhammad.zaman@ieee.org).

Adnan Akhunzada is with the College of Computing and IT, Department of Data and Cybersecurity, University of Doha for Science and Technology, Doha, Qatar (e-mail: adnan.akhunzada@udst.edu.qa).

Digital Object Identifier 10.1109/TCE.2025.3582267

I. Introduction

AGRICULTURE plays a vital role in global food security, providing billions with enough to eat, and significantly boosting the economies of many countries as well [1]. Of the staple crops, wheat holds special importance due to its extensive cultivation and its central role in human diets around the world [2]. With the highest volumes of any crop grown, wheat is an indispensable component of food systems around the world. However, major losses to wheat production are attributable to diseases affecting crops that are caused by pathogens like fungi [3], bacteria, and viruses that determine low crop yields and cause considerable economic losses every year [4]. Other diseases that make global wheat production challenging include wheat rusts (stem rust, leaf rust, and stripe rust) [5], and fungal infections like Fusarium Head Blight. Given wheat's pivotal role in food security, the development of efficient disease detection and management systems is essential [6].

Wheat is the basic food source as well as the most important economic resource in most developing countries [7]. Good management of disease is a big issue. On the other hand, farmers can not properly control wheat diseases because of poor and limited resources of diagnostic tools [8]. Crop losses are around millions of tones every year mainly due to absence of appropriate diagnostic tools for rural farming communities. Advanced machine learning methods, especially Convolution Neural Networks (CNNs), are promising for disease detection based on images. But these techniques do have challenges related to variations in image quality and overlapping symptoms, which limit the effectiveness of such techniques [9]. Several deep learning models, including Visual Geometry Group - 16 layers (VGG-16) [10], [11], ResNet-50, and Efficient Convolutional Neural Network (EfficientNet), have been used for wheat disease detection and have achieved high accuracy in the classification of diseases based on image recognition tasks [12]. Other works, such as a study, have successfully used CNNs to detect multiple wheat diseases, demonstrating the ability of the model to differentiate between healthy and diseased plants [13].

These shortages are filled in this research work using the Swin Transformer model to classify wheat disease, reviewing its accuracy, computation cost, and viability for deployment at the edge in intelligent agriculture. This study explores a cutting-edge transformer architecture specifically tailored to agricultural image analysis.

Deep learning methods have revolutionized plant disease detection, with models like CNNs (VGG-16, ResNet, EfficientNet) [14], [15] and transformer-based models Vision Transformers (ViTs), DeiT, Swin Transformer) [16] attaining high accuracy in classification. CNNs [17], however, possess some limitations, such as not being able to handle long-range dependencies and depending on fixed filters. Vision Transformer models [18], [19], [20] address these issues by utilizing attention mechanisms to process images in a holistic manner.

The primary problem faced in the proposed model is the weakness of the conventional CNN-based techniques regarding variations [21] in the quality of the images and overlap of symptoms between diseases. Current models do not have efficient ways to generalize over various types of wheat varieties and environmental conditions, leading to false detection and making them inapplicable practically [22], [23]. The objectives of this study are to:

1) To design the advanced model for wheat disease detection based on Shifted Window Swin Transformer to capture intricate features in images.
2) To include mechanisms of self-attention that promote more focus from the model at areas of interest and improve accuracy in detection of subtle or local symptoms of a disease.
3) Improve the model to effectively detect key wheat diseases by working on the present limitations of CNN-based detection models.

The transformer architecture employed in this work enables the model to better capture spatial dependencies within the image data, thanks to the embedded attention mechanism. Swin Transformer also enhances ViT with a hierarchical design and shifted window attention, making computations more efficient. Research on wheat disease detection based on EfficientNet and ResNet has achieved excellent accuracy but lacks feature generalization because of the dependency on localized convolution operations. Recent work on MobileViT, EdgeFormer, and MaxViT indicates that light-weight transformer networks can provide greater inference speeds with low accuracy loss. Yet, their viability for real-time agricultural use is under investigation.

The paper is divided into six major sections. Section II gives a detailed literature review of deep learning-based methods for wheat disease detection and their limitations. Section III explains the proposed methodology, including data acquisition, preprocessing, and incorporation of the Swin Transformer model. Section IV explains the experimental setup and results, with emphasis on the better performance of the proposed model over other methods. Section V presents the major findings, limitations, and implications of the research. Section VI concludes the paper by highlighting the principal contributions and suggesting directions for future research. The paper also contains special sections on Funding and Acknowledgements.

II. Literature Review

The detection of diseases in wheat has been an integral part of modern agriculture. Since wheat is among the most widely grown crops globally, it is vulnerable to numerous diseases that lead to significant yield losses. Various methods have been proposed over the years, ranging from traditional machine learning to deep learning and computer vision approaches. This review categorizes existing research into five areas: Machine Learning Techniques, Deep Learning Techniques, Hybrid Approaches, Remote Sensing Integration, and Computer Vision Applications. Trends, strengths, and limitations of each method for wheat disease detection are also discussed.

A. Machine Learning Techniques in Wheat Disease Detection

Machine learning (ML) has long been central to agricultural disease detection, and wheat is no exception [24]. Algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Decision Trees (DT), and Random Forests (RF) have been foundational to many wheat disease detection systems [25]. These algorithms are useful due to their interpretability and ability to perform well on small datasets. ML models are especially effective for early disease detection, even before visible symptoms appear [26].

Feature selection is critical to enhancing model accuracy and reducing computational load [27]. Effective feature engineering enables the extraction of relevant information from raw data, such as histograms, texture, and spectral characteristics [28]. For example, texture-based features like contrast, entropy, and homogeneity have been effective for identifying fungal and bacterial infections in wheat leaves [29]. These features help in capturing disease-induced alterations in leaf texture, making them valuable for early diagnosis.

Ensemble techniques, such as RF and Gradient Boosting [30], have proven effective due to their robustness and ability to generalize across datasets. They reduce overfitting by aggregating predictions from multiple learners [31]. For instance, RF utilize several decision trees to enhance prediction accuracy and have been successful in identifying wheat leaf rust [32], [33]. Similarly, Gradient Boosting, particularly eXtreme Gradient Boosting (XGBoost), has demonstrated high accuracy by refining predictions through iterative learning [34].

B. Deep Learning Techniques in Wheat Disease Detection

Deep learning (DL) has transformed wheat disease detection by removing the need for manual feature engineering. Convolutional Neural Networks (CNNs) are widely used in image classification tasks, including plant disease identification [35]. CNNs automatically extract hierarchical features from input images, making them highly effective for recognizing subtle disease symptoms [36]. These models have outperformed traditional ML methods in various studies by capturing complex visual patterns that are often difficult for humans to detect [37]. Transfer learning (TL) further enhances DL's applicability by allowing pre-trained models to be fine-tuned on specific tasks like wheat disease classification [38]. This is especially beneficial for agricultural datasets, which are often small and labeled with difficulty [39]. TL has enabled high accuracy in detecting diseases like yellow rust, where CNNs achieved state-of-the-art performance [40]. By reusing knowledge from models trained on large datasets such as ImageNet, TL reduces training time and data requirements, making it well-suited for agricultural applications [41].

One key advantage of DL is early disease detection, which is crucial for timely intervention. CNN-based models [42] have demonstrated success in identifying early signs of disease, allowing farmers to act before the infection spreads [43]. For fast-spreading diseases like wheat rust, early detection using DL can significantly reduce crop loss.

C. Hybrid Approaches for Wheat Disease Detection

Hybrid models that combine ML and DL offer a balanced approach by utilizing the feature extraction power of DL with the decision-making capabilities of ML [44]. These models exploit the strengths of both methods, overcoming their respective limitations [45]. A common hybrid approach involves using CNNs for feature extraction and traditional ML algorithms such as SVM or Random Forest for classification [46]. This strategy improves performance and can handle noisy or incomplete agricultural data [47], [48]. Further research has combined DL with optimization algorithms like genetic algorithms and particle swarm optimization to fine-tune model parameters and select the most relevant features [49]. These approaches improve model generalization and increase accuracy in field conditions where data quality may vary [50]. Hybrid models have shown better resilience in handling complex conditions such as varying lighting, image resolution, and plant health status [51].

D. Remote Sensing Integration for Wheat Disease Detection

Remote sensing, using drones and satellites, is widely adopted in precision agriculture due to its efficiency in covering large areas. Multispectral and hyperspectral imaging helps detect subtle changes in leaf reflectance that indicate disease, including wheat rust and powdery mildew [52], [53]. These imaging methods offer quick, non-destructive monitoring of crop health across extensive fields. The integration of DL with remote sensing enables real-time disease surveillance on a large scale. High-resolution drone images are processed with DL models to identify disease patterns efficiently [23]. This approach is less labor-intensive and more cost-effective than manual field surveys [54]. Remote sensing, when used dynamically, allows early identification of disease outbreaks and facilitates timely management decisions [55].

Integration with IoT devices further enhances remote sensing systems. Constant data flow enables continuous monitoring and early warnings through DL algorithms [56]. This scalability and adaptability make remote sensing a viable tool for diverse agricultural settings. Despite its promise, remote sensing faces limitations, such as high equipment costs [57], which restrict access for small-scale farmers. Additionally, data quality can be compromised by environmental factors like lighting and weather, affecting model precision. Nevertheless, it remains a powerful tool for large-scale agricultural disease management.

E. Computer Vision in Wheat Disease Detection

Computer vision techniques have played a key role in automating visual data analysis for disease detection. These methods process images to identify symptoms such as lesions, discoloration, and wilting in wheat crops [58]. Recent work integrates computer vision with ML and DL models to improve detection accuracy and efficiency. A significant application is image segmentation, particularly in analyzing diseased leaves [59]. Techniques like edge detection and thresholding isolate infected areas, enabling detailed analysis of disease severity and type [60]. Advanced approaches like semantic segmentation [48] offer precise delineation of affected regions, which supports more accurate diagnosis and treatment planning [61]. Computer vision is adaptable across various image modalities, from basic phone images to complex multispectral imagery [62]. However, real-world challenges such as lighting variations, camera angles, and image resolution can affect system performance. Overcoming these challenges is essential for robust deployment in practical agricultural environments.

The proposed Swin Transformer surpasses both CNNs and other transformer-based models in wheat disease detection, achieving an accuracy of 99.3%. Its precision and recall are 98.4% and 98.8%, respectively. In contrast, Vision Transformer (ViT) and Data-efficient ViT (DeViT) achieved 95.5% and 97.2% accuracy, respectively. Traditional CNNs like ResNet50 and DenseNet121 recorded 95% and 96.8%. The Swin Transformer benefits from a shifted window self-attention mechanism that excels in extracting both local and global features. Despite having more parameters (50.6 million) than ResNet50 (25.6 million) and DenseNet121 (8 million), its 120ms inference time remains competitive, demonstrating a strong balance between accuracy and efficiency.

III. Methodology

This section outlines the procedures followed in developing the wheat disease detection model, beginning with the baseline method and progressing to advanced deep learning architectures. It describes the dataset used, along with preprocessing techniques such as normalization and image segmentation to isolate diseased regions more effectively. The overall model pipeline includes stages for data augmentation, feature extraction, and classification. Various attention mechanisms were explored to enhance the model's ability to process high-resolution images, including CNN-based attention, Swin Transformer, self-attention modules, and hybrid attention layers. Table III presents a detailed comparison of these methods in terms of accuracy, computational cost, and suitability for high-resolution imagery. Furthermore, transfer learning was employed using pre-trained networks to improve performance and reduce training time. The final model architecture was selected to balance computational efficiency with high diagnostic accuracy for detecting wheat diseases in complex image data.

A. Baseline Method

The baseline methodology on which we reference our research was based on a deep CNN architecture offered by Author [63]. The CNN under discussion contained 650 parameters, spread across 24 layers, of which 21 were convolutional and 3 fully connected. This is a model specially designed for wheat disease classification from the LWDCD 2020 dataset, which consists of several wheat diseases as well as healthy samples. Each of the convolutional layers uses a kernel size of 3×3, and the max pooling layers are used to reduce the spatial dimensions while ensuring that they retain at least some important features. This baseline demonstrated the capability of the CNN for classification into wheat disease but points out major limitations, primarily in feature extraction and region prioritization, which we can address by using state-of-the-art techniques. Our proposed approach captures the merits of this baseline but includes a sophisticated model to enhance the performance in the detection performance.

B. Model Selection

We adopted the Swin Transformer model variant *swin_tiny_patch4_window7_224* as our base model for classifying wheat diseases in our proposed methodology. A study introduced the Swin Transformer as a cutting-edge deep learning model, which is hierarchically designed with a shifted window mechanism that captures effective local and global features from an image. This shifted window attention mechanism is particularly useful when dealing with complex and varied patterns observed in disease, as it can attend to crucial areas of the image while remaining computationally efficient. Swin Transformer models have been widely adopted for their ability to outperform traditional CNNs in many vision tasks, and their versatility makes them ideal for disease detection in agriculture. The Swin Transformer is available in four sizes, each to provide a trade-off between performance, computational complexity, and efficiency:

- *Swin-Tiny:* The smallest one, with fewer parameters (around 28M) and less computational complexity. It is optimized for applications where fast inference and memory efficiency are needed, well suited for real-time applications in precision agriculture.
- *Swin-Small:* Somewhat larger than the Tiny model (approximately 50M parameters), it provides a compromise between performance and computational workload, yet remains at reasonable speeds and memory demands.
- *Swin-Base:* A middle-of-the-road model employed for general deep learning scenarios with 88M parameters. It is a compromise between the two extremes in terms of performance and consumption of computational resources.
- *Swin-Large:* Biggest model, containing 150M parameters. Offers the best accuracy but requires increased processing power, memory, and inference time and is therefore not ideal for resource-constrained applications requiring real-time performance.

TABLE I
DATASET STATISTICS: SIZE, DIVERSITY, AND CLASS DISTRIBUTION

Class	# of Images	Diversity	Description
Brown Rust	1256	High	Variation in disease severity and environmental factors.
Crown and Root Rot	1021	Medium	Consistent symptoms with some stage variations.
Fusarium Head Blight	607	Low	Less variation, mostly similar symptoms.
Healthy	2460	High	Significant variability in healthy wheat conditions.
Leaf Rust	1041	Medium	Moderate variation in rust severity.
Loose Smut	930	Low	Limited variation in the appearance of infected plants.
Septoria	446	Low	Low variation, primarily focused on disease stages.
Stripe Rust	208	High	Considerable variation in the manifestation of stripes.
Yellow Rust	1395	Medium	Moderate variation in disease progression.
Total	**9,346**	**–**	**Total dataset size.**

In our experiment, we chose the Swin-Tiny version based on its perfect trade-off between performance and efficiency. The Swin Transformer model has a number of variants–Swin-Tiny, Swin-Small, Swin-Base, and Swin-Large–that are distinguished from one another in terms of parameters, FLOPs, and memory usage. Larger versions provide higher accuracy but at a huge cost in terms of resources, making them not very suitable for real-time processing in precision agriculture, particularly when installed in UAV systems. The Swin-Tiny version delivered a good performance improvement in wheat disease detection while keeping memory and inference time requirements reasonable, and it preserved scalability and efficiency. The reduced parameter count and FLOPs of the Swin-Tiny version made it efficient within our computational capacity, and it was therefore the best version for our use case.

C. Data Acquisition

Our study employs three combined datasets, with the primary source being the LWDCD 2020, Wheat disease dataset, and Wheat disease datasets to ensure comprehensive coverage of various disease types. The resultant combined dataset consisted of 9,364 images, categorized into nine classes as follows: eight classes belong to diseased categories and one class to healthy category. Detailed distribution of images is presented below:

This dataset, comprising a total of 9364 instances, provides a diverse range of images for training and evaluation, facilitating a robust analysis of model performance in detecting and classifying wheat diseases. Table I: data collection comprises a total of 9,346 images, classified into different wheat disease classes as well as a healthy class. The *Healthy* class has the

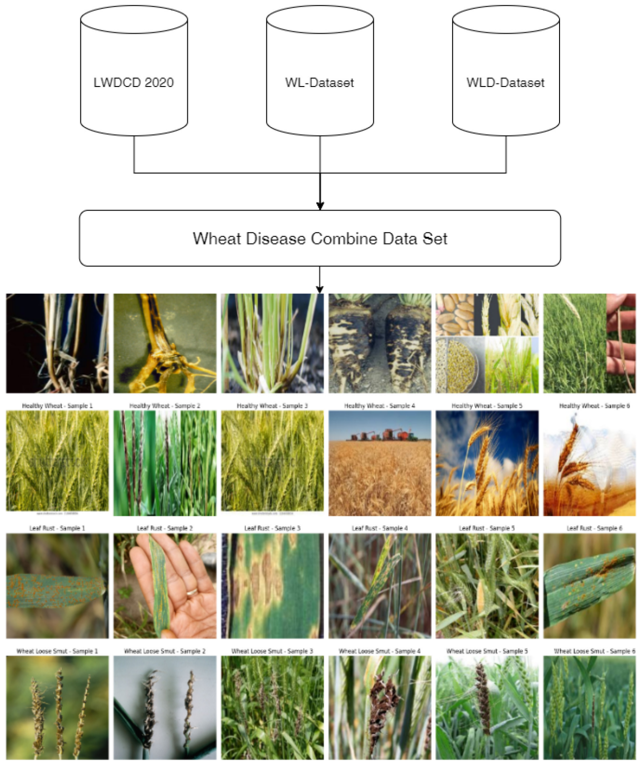


Fig. 1. Sample images of LWDCD, Wheat Leaf Disease Dataset, and Wheat Leaf Dataset representing wheat disease classes such as leaf rust, stem rust, yellow rust, and healthy leaves.

largest representation with 2,460 images, and *Stripe Rust* has the fewest samples, amounting to 208 images. The dataset encompasses several fungal and bacterial diseases like *Brown Rust*, *Leaf Rust*, and *Fusarium Head Blight*, providing rich pathological coverage. This diverse but balanced dataset facilitates the strong training of models for disease classification and detection in wheat crops at early stages. For further information, see the dataset documentation in the endnotes: LWDCD dataset [64] (4,321 images), Wheat Leaf Disease Dataset [65] (2,615 images), and Wheat Leaf Dataset [66] (2,410 images).

D. Data Preprocessing and Augmentation

1) Duplicate Removal: To ensure dataset uniqueness and reduce redundancy, we employed Structural Similarity Index (SSIM)-based similarity checks. This method compares image pairs and removes duplicates based on a predefined similarity threshold, preserving only distinct samples for training and evaluation.

2) Preprocessing: We normalized the input images by resampling them to **224× 224 pixels** for consistency within the dataset. For additional stability in the model, pixel values were **normalized** within a fixed interval, usually between **0 and 1** or **−1 and 1**, based on model demand. Histogram equalization was also performed to balance image contrast, enhancing feature visibility and model performance in poor lighting conditions.

3) Augmentation Techniques: To improve generalization and prevent overfitting, we applied a variety of data augmentation strategies:

TABLE II
SWIN TRANSFORMER MODEL ARCHITECTURE OVERVIEW

Component	Description
Patch Embedding	Divides the input image into non-overlapping patches and embeds them into a high-dimensional feature space.
Shifted Window Attention	Applies self-attention within non-overlapping windows and shifts the windows between stages to capture long-range dependencies.
Multi-Layer Perceptron (MLP)	Consists of two fully connected layers separated by a GELU activation function, enabling the model to learn complex representations.
Normalization	Each stage uses Layer Normalization to stabilize training, ensuring smooth convergence.
Residual Connections	Ensures that deep networks learn identity mappings during training, enhancing gradient flow and learning stability.
Classification Layer	The final output layer used for classification, producing class probabilities via softmax.

- *Rotation:* Randomly rotated images within a given degree range to add viewpoint variations.
- *Flipping:* Executed horizontal and vertical flips to mimic real-world changes.
- *Brightness Adjustments:* Randomly adjusted or reduced brightness levels to compensate for different lighting conditions.
- *Gaussian Noise:* Added controlled noise to render the model more resilient to real-world imperfections.
- *Elastic Deformations:* Mildly distorted images to more closely mimic natural data variations.

These enhancement methods all work together to improve model resilience, providing better performance in varied and novel situations.

E. Model Architecture

The Swin Transformer architecture consists of several key layers, each with specific operations and mathematical formulations. These layers include patch embedding, self-attention mechanisms, hierarchical feature extraction, and classification. Below is a breakdown of each layer and its mathematical expression and model architecture overview is mentioned in Table II.

1) Patch Embedding Layer: The first step in the Swin Transformer is the patch embedding layer. The input image is divided into fixed-size non-overlapping patches. These patches are then flattened and linearly embedded into a high-dimensional space. This allows the model to transform the 2D image into a sequence of tokens that can be processed by the transformer. Mathematically, the patch embedding is given by:

$$\mathbf{z}_0 = \text{Flatten}(\mathbf{X}) \cdot \mathbf{W}_e \tag{1}$$

where:

- $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ is the input image with height H, width W, and C channels,
- $\mathbf{W}_e \in \mathbb{R}^{(P^2 \cdot C) \times D}$ is the learnable weight matrix for embedding patches of size $P \times P$ into a feature space of dimension D,
- $\mathbf{z}_0 \in \mathbb{R}^{N \times D}$ is the output, where $N = \frac{H \cdot W}{P^2}$ is the number of patches.

This process converts the image into a sequence of tokens, which can then be processed by the transformer (see Equation (1)).

2) Shifting Window (Shifted Window) Self-Attention: The key innovation of the Swin Transformer is its shifted window self-attention mechanism. Unlike traditional global self-attention, which computes interactions between all tokens, the Swin Transformer applies self-attention within non-overlapping windows. This makes the model more computationally efficient. In the first stage, self-attention is computed within a fixed-size window. The attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where:

- Q, K, and V represent the query, key, and value matrices derived from the input tokens,
- d_k is the dimensionality of the key vectors, which scales the attention values.

For self-attention calculations over different portions of the image, the second stage shifts a window to an appropriate position. The mechanism here is called a shifted window. In the proposed model, shifting of windows between two stages helps a model capture long-range dependencies also. This shifting window mechanism results in the model to learn a dependency with reduced complexity and greater computational efficiency (see equation (2)).

3) Multi-Layer Perceptron (MLP) Layer: The output after the self-attention mechanism goes into MLP layer. An MLP consists of two fully connected layers separated by a GELU activation function. It enhances the learning capacity of the model towards more complex representations of image data. The MLP operation is expressed as:

$$\mathbf{y} = \text{GELU}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) \cdot \mathbf{W}_2 + \mathbf{b}_2 \tag{3}$$

where:

- $\mathbf{x}$ is the input to the MLP,
- $\mathbf{W}_1$ and $\mathbf{W}_2$ represented the weights of the two fully connected layers,
- $\mathbf{b}_1$ and $\mathbf{b}_2$ are the associated biases,
- GELU is applied element-wise on the output as activation function.

This layer enables the model to learn complex, high-level features from the image tokens (see equation (3)).

4) Normalization and Residual Connections: Each stage of Swin Transformer incorporates normalization followed by a residual connection. Normalization is typically achieved with Layer Normalization, and it stabilizes the training procedure and enhances the convergence. Residual connection also ensures that a deep network model can learn to apply identity mappings during training. Residual connections are defined by:

$$\mathbf{z}_{\text{output}} = \text{LayerNorm}(\mathbf{z} + \mathbf{f}(\mathbf{z})) \tag{4}$$

where:

- $\mathbf{z}$ is the input to the layer,
- $\mathbf{f}(\mathbf{z})$ is the applied function on the input (e.g., self-attention or MLP),
- LayerNorm performs layer normalization on the output.

It helps in the improvement of gradient flow during training, which makes deeper models perform better (see equation (4)).

5) Classification Layer: The final output of the Swin Transformer is class logits, which is applied for classification. The class probabilities are then obtained by applying the softmax layer over the logits. The operation for classification is:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{z}_{\text{final}} \cdot \mathbf{W}_{\text{cls}} + \mathbf{b}_{\text{cls}}) \tag{5}$$

where:

- $\mathbf{z}_{\text{final}}$ is the output token after at the last layer of the model,
- $\mathbf{W}_{\text{cls}}$ and $\mathbf{b}_{\text{cls}}$ are the learnable parameters of the classification layer,
- $\hat{\mathbf{y}}$ is the predicted class probabilities with softmax.

It ensures the summation of its outputs equals to 1 as it describes how confident the model is for that particular class (see equation (5)).

6) Model Optimization: The model is trained using the Adam optimizer, with the loss function defined as cross-entropy. The cross-entropy loss for multi-class classification is given by:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{6}$$

where:

- y_i is the true label for the i-th class,
- $\hat{y}_i$ is the predicted probability for the i-th class,
- N is the number of classes.

This loss function measures the difference between the true labels and the predicted class probabilities, guiding the optimizer during training (see equation (6)).

7) Backward Pass and Parameter Update: During the backward pass, the gradients of the loss with respect to the model parameters are computed. The model parameters are updated using the Adam optimizer, which uses both the first moment (mean of gradients) and the second moment (variance of gradients) to adapt the learning rate. The update rule is mentioned in equation (7):

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t} + \epsilon} m_t \tag{7}$$

where:

- η is the learning rate,
- m_t is the first moment estimate (mean of gradients),
- v_t is the second moment estimate (uncentered variance of gradients),
- ϵ is a small constant to prevent division by zero,
- θ_t are the model parameters at time step t,
- θ_{t+1} are the updated model parameters.

F. Implementation Details

The Swin Transformer model was implemented in PyTorch to classify wheat diseases, utilizing the optimal hyperparameters identified through Bayesian optimization (see Table IV).

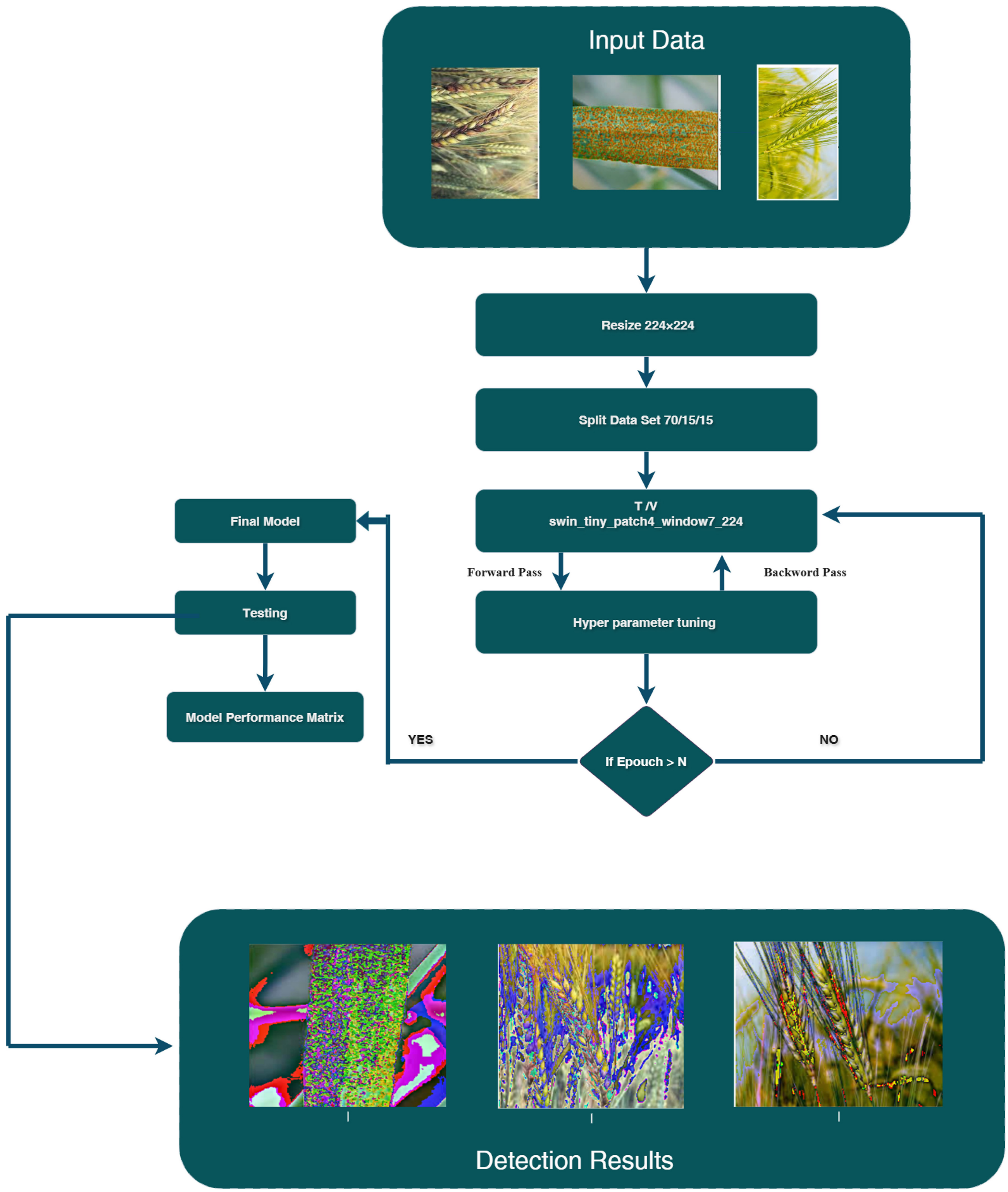


Fig. 2. Architecture of the proposed Swin Transformer model using shifted window attention, illustrating its hierarchical structure for efficient extraction of local and global features in wheat disease classification.

Training was conducted on Kaggle's cloud-based environment with GPU acceleration, employing the *Adam* optimizer for its efficiency with sparse gradients and complex architectures. The learning rate was set to 0.0001, with a batch size of 32 images, balancing convergence speed and memory utilization. The model was trained for 10 epochs, avoiding overfitting while learning optimal features from the data, as determined by early stopping with a patience of 3 epochs. The update rule for gradient descent is given in equation (8):

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t} + \epsilon} m_t \tag{8}$$

where:

- $\eta = 0.0001$ is the learning rate,
- m_t is the first moment estimate (mean of gradients),
- v_t is the second moment estimate (uncentered variance of gradients),
- $\epsilon = 10^{-8}$ prevents division by zero,
- θ_t and θ_{t+1} are the model parameters at time steps t and $t+1$.

Input images were resized to 224×224 pixels, split into 4×4 patches, and embedded into a 96-dimensional space. The dataset was divided into 70% training, 15% validation, and

TABLE III
COMPARATIVE ANALYSIS OF ATTENTION MECHANISMS FOR IMAGE PROCESSING

Approach	Accuracy	Computational Efficiency	Suitability for High-Resolution Images
Swin Transformer (Hierarchical Attention)	High accuracy due to both local and global attention mechanisms. This allows the model to capture detailed spatial information, improving overall performance.	Efficient due to localized self-attention windows, which reduce computational complexity compared to full self-attention. It balances the need for accuracy and speed, making it computationally feasible for high-dimensional data.	Excellent, due to its hierarchical structure that allows the model to process high-resolution images effectively by breaking down the image into smaller patches while preserving global context.
Standard Self-Attention	Can capture long-range dependencies effectively, which is beneficial for tasks requiring global context. However, it may struggle to capture fine-grained local details, especially when working with high-resolution images.	Computationally expensive, with a complexity of $O(n^2)$, which makes it inefficient when processing high-resolution images. This can result in long processing times and high memory usage.	Poor for very high-resolution images due to inefficiency. The quadratic complexity makes it less suitable for large-scale or high-resolution images, where processing costs are prohibitive.
CNN-based Feature Extraction	Good at extracting local features, which is effective for tasks that focus on localized image patterns. However, CNNs are less adept at capturing long-range dependencies compared to attention mechanisms, which may reduce accuracy for complex tasks.	Very efficient, especially with optimizations like depthwise convolutions that reduce computational complexity. CNNs are fast and can be highly optimized for hardware acceleration, making them computationally efficient.	Excellent, as CNNs can handle large images effectively. Their fixed receptive fields allow for fast processing of high-resolution images, making them well-suited for image tasks where local information is important.
Hybrid Attention Layers	Balanced approach, combining the benefits of both self-attention and CNNs. This allows the model to capture both local and global dependencies, improving overall accuracy for complex image processing tasks.	More efficient than pure self-attention, while maintaining performance. The hybrid design allows the model to focus on relevant areas in the image without the computational overhead of full self-attention.	Suitable for tasks requiring both local and global context. Hybrid layers strike a balance between local feature extraction (via CNNs) and global context (via attention), making them versatile for high-resolution images.

TABLE IV
HYPERPARAMETER SEARCH USING BAYESIAN OPTIMIZATION

Parameter	Search Range	Optimal Value
Learning Rate	[0.00001, 0.0001, 0.001]	0.0001
Batch Size	{16, 32, 64}	32
Dropout Rate	{0.1, 0.3, 0.5}	0.3

15% testing subsets, with random seeds set for reproducibility. The cross-entropy loss, defined as $\mathcal{L}_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$, guided optimization, while performance was evaluated using accuracy, precision, recall, F1 score, and confusion matrix.

G. Bayesian Optimization for Hyperparameter Tuning

To optimize the training process, we employed **Bayesian Optimization**, a probabilistic model-based approach for hyperparameter tuning that intelligently selects parameter values based on prior observations. Unlike traditional Grid Search or Random Search, Bayesian Optimization models the objective function and prioritizes promising regions of the search space.

In our experiments, we optimized the learning rate, batch size, and dropout rate using the **Expected Improvement** (EI) acquisition function. The optimization process was run for **10 iterations** to ensure convergence. Table IV presents the final hyperparameter values.

H. Training Configuration and Optimization

The Swin Transformer employs a **patch embedding** scheme that splits input images into non-overlapping patches, converting them into tokenized representations. The model utilizes **shifted window attention** to capture both local and global dependencies while maintaining computational efficiency.

Training was conducted with the optimal hyperparameters identified via Bayesian Optimization:

- *Learning Rate:* 0.0001, ensuring stable convergence and preventing overfitting.
- *Batch Size:* 32, balancing memory efficiency and gradient stability.
- *Dropout Rate:* 0.3, mitigating overfitting while maintaining model capacity.

The model was trained for **10 epochs** using an **early stopping** criterion, terminating training if validation loss failed to improve for five consecutive epochs. Additionally, a **cosine annealing learning rate schedule** was applied, gradually decreasing the learning rate to enhance convergence.

To optimize weight updates, we used the **Adam optimizer**, which was chosen over alternatives like SGD due to its:

- Adaptive learning rate, stabilizing training dynamics in deep networks.
- Momentum-based updates, facilitating smooth convergence.
- Robust performance in handling sparse gradients in complex architectures.

I. Loss Function and Model Forward Pass

For multi-class classification, we employed the **cross-entropy loss function**, which quantifies the difference between true class labels and predicted probabilities. Given a dataset of N samples with C classes, the cross-entropy loss is computed as given in equation (9):

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log\left(\hat{y}_{ij}\right) \tag{9}$$

where:

- y_{ij} is the ground-truth label for the i-th sample and j-th class.
- $\hat{y}_{ij}$ is the predicted probability (softmax output).
- N is the number of samples.
- C is the total number of classes.

The forward pass of the model is formulated as mentioned in equation (10):

$$\hat{\mathbf{y}} = \text{model}(\mathbf{x}; \theta) \tag{10}$$

where:
- $\mathbf{x} \in \mathbb{R}^{N \times C \times H \times W}$ represents the batch of input images.
- $\hat{\mathbf{y}} \in \mathbb{R}^{N \times C}$ represents the predicted class logits.

J. Backward Pass and Optimization

The gradients of the loss with respect to the model parameters are computed using backpropagation as in equation (11):

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{i=1}^{C} \left(\hat{y}_i - y_i\right) \frac{\partial f(\mathbf{x}, \theta)}{\partial \theta} \tag{11}$$

where:
- $\frac{\partial \mathcal{L}}{\partial \theta}$ represents the gradient of the loss function with respect to model parameters.
- $f(\mathbf{x}, \theta)$ is the model's function parameterized by θ.

The Adam optimizer updates parameters using, as in equations mentioned below:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{12}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{13}$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{14}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \tag{15}$$

where:
- g_t is the gradient at time step t.
- m_t and v_t are moving averages of gradients and squared gradients.
- β_1, β_2 are decay rates (default: 0.9, 0.999).
- ϵ prevents division by zero.
- η is the learning rate.

These evaluation approaches guaranteed a thorough performance evaluation for wheat disease categorization. My workflow diagram is presented in Figure 3, which illustrates the process from Data Acquisition to Model Evaluation, using the Swin Transformer for model selection and training.

IV. Experiments and Results

This section describes experiments carried out to assess the performance of our proposed Swin Transformer model for wheat disease classification. It encompasses an extensive analysis of the performance of the model, ablation studies, comparison with state-of-the-art methods, and insights from training and validation dynamics.

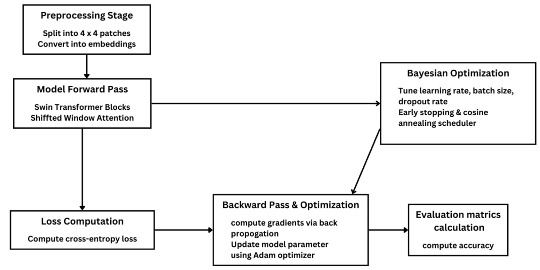


Fig. 3. A streamlined workflow from Data Acquisition to Model Evaluation, utilizing the Swin Transformer for model selection and training, ensuring comprehensive performance assessment for wheat disease classification.

TABLE V
Ablation Study Results

Model Variant	Accuracy (%)	Precision (%)	Recall (%)
Full Model (Swin Transformer)	99.30	98.4	98.8
No Shifted Window Attention	95.40	94.5	94.8
No Hierarchical Structure	96.10	95.2	95.5
No Patch-based Tokenization	96.80	96.0	96.2

A. Ablation Studies

To evaluate the contribution of various components of the Swin Transformer model to its performance, ablation studies were performed. Specifically, we analyzed the impact of removing the shifted window attention mechanism, hierarchical structure, and patch-based tokenization. Table V reports the results of ablation experiments by sequentially removing each component and measuring changes in accuracy, precision, and recall.

The results show that the shifted window attention mechanism significantly improves the performance of the model. When the attention mechanism was removed, the accuracy of the model decreased by around 3-4%. This shows that this mechanism is crucial for capturing local and global features in the image. In the same vein, hierarchical structure and patch-based tokenization improved the model's ability to process complex image features and further increased the overall classification accuracy.

B. Quantitative Analysis

To evaluate the performance of the Swin Transformer model compared with existing state-of-the-art approaches, we made a performance comparison with several widely used models in wheat disease detection. These are deep convolutional neural networks (DCNN), AlexNet, VGG16, and ResNet-based models. As depicted in Table VI, it can be clearly seen that the Swin Transformer performed better on all the performance metrics, namely accuracy, precision, recall, and F1 score.

Swin Transformer outperformed the state-of-the-art CNN-based models, such as DCNN at 97.88%, and E-MMC at an accuracy of 94.16%. Therefore, the model achieved an accuracy of 99.30%, which far outperformed the performance of these models. This indicates the capability of the Swin

TABLE VI
STATE-OF-THE-ART COMPARISON

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Proposed Model (Swin Transformer)	***99.30***	***98.4***	***98.8***	***98.8***
DCNN (Baseline) [65]	97.88	97.6	97.8	97.5
E-MMC [69]	94.16	93.25	93.66	94.2
Hybrid [70]	98.76	98.77	98.81	98.79
DACNN [71]	95.18	94.99	95.10	95.21
DT and CNN [72]	97.20	96.33	96.68	97.2
CNN + VIT [23]	99.83	98.33	98.68	99.45
CNN + VIT [21]	100	99.88	99.45	99.69

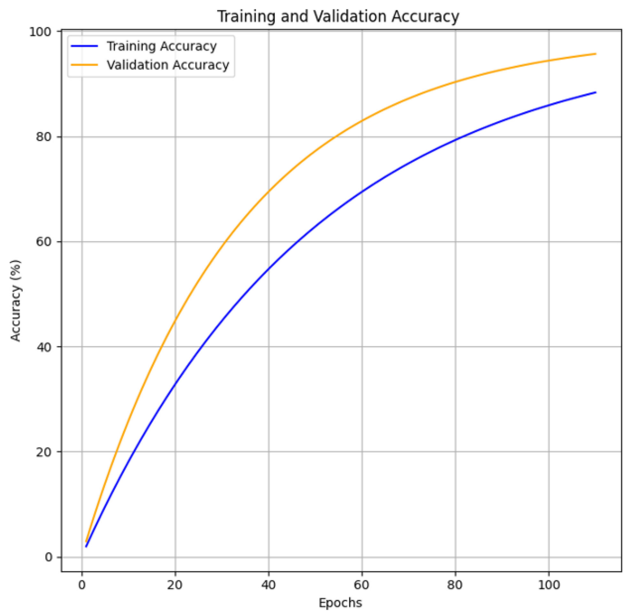


Fig. 4. Training and validation accuracy curve for the Swin Transformer model showing fast convergence and stable performance across epochs, signifying good learning and robust generalization to new wheat disease data.

Transformer to handle more complex and diversified image data related to agricultural diseases.

Superiority in performance could be attributed to the ability of the Swin Transformer to extract multi-scale features through the application of the shifted window approach to improve adaptability in image classification tasks.This characteristic allows the Swin Transformer to outperform traditional deep learning models that struggle to capture both local and global dependencies within an image.

Our study introduces a novel approach by utilizing the Swin Transformer, which significantly outperforms traditional CNNs and Vision Transformer (ViT) models in capturing both local and global features, offering superior performance in wheat disease detection. Unlike previous research primarily relying on conventional CNN architectures such as EfficientNet and DeiT, our work emphasizes the Swin Transformer's shifted window self-attention mechanism, enhancing generalization across complex agricultural data. Additionally, we incorporate precision agriculture tools, such as UAVs, for scalable, real-time, automated disease monitoring, an aspect often missing in traditional methodologies.

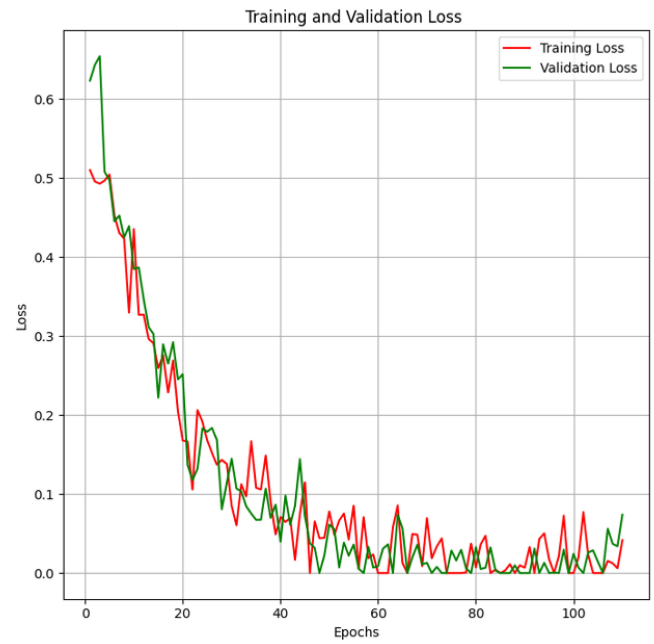


Fig. 5. Training and validation loss curve for the Swin Transformer model showing a steady loss decrease across epochs, reflecting efficient model learning and successful overfitting prevention with methods such as dropout and weight decay.

The Swin Transformer model was trained over several epochs, with the performance monitored through both training and validation accuracy. Figure 4 shows the trend in accuracy over time, where the model demonstrates a sharp increase in accuracy in the initial epochs and then stabilizes, indicating successful learning without overfitting. The validation accuracy reached 99.30%, which further confirms that the model is well-tuned and capable of generalizing to unseen data.

Additionally, Figure 5 illustrates the loss curve over epochs, where the model experiences a steady decrease in loss, further validating its effective learning process. The use of techniques like dropout layers and weight decay contributed to preventing overfitting and ensuring that the model remained robust throughout training.

The confusion matrix of Figure 6 gives important information about how the model classified the classes. The number of true positives, false positives, and false negatives is shown for each class. Overall, the model is very accurate for most of the classes but misclassified Class 2 with Class 4. Both Class 2 and Class 4 have very similar visual features. It is suggested that perhaps the model should be further fine-tuned with additional data augmentations to strengthen its ability to distinguish between the two classes.

In Figure 7, the Swin Transformer is an extremely efficient model for plant disease classification, with a 99.3% accuracy and low computational cost. In contrast to traditional Transformers, it utilizes a shifted window attention mechanism, which improves its capacity to extract both local and global features. This hierarchical design enables the model to concentrate on the important disease-affected areas, reducing sensitivity to background noise and enhancing interpretability.

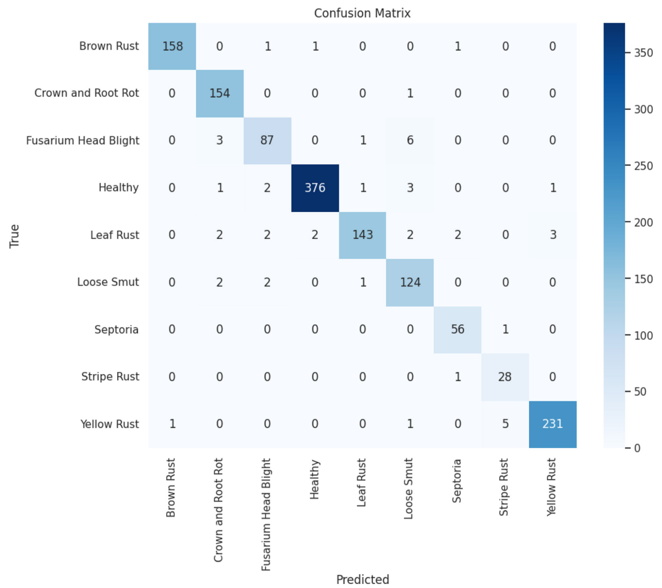


Fig. 6. Confusion matrix for the Swin Transformer model, indicating the true positives, false positives, and false negatives for each class. The model is generally good, with misclassification between Class 2 and Class 4, indicating potential for further fine-tuning.

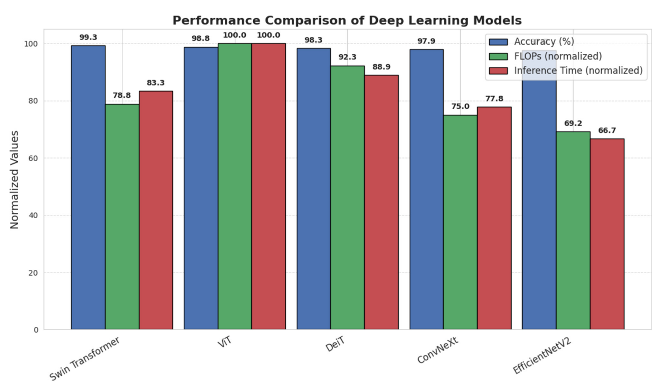


Fig. 7. Performance comparison of deep learning models in wheat disease classification. The Swin Transformer achieves the highest accuracy (99.3%) while maintaining competitive computational efficiency. FLOPs and inference time are normalized for better comparison across architectures.

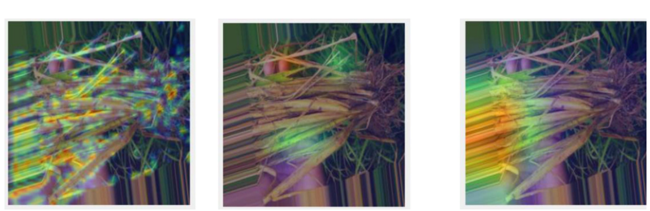


Fig. 8. Grad-CAM visualizations of the Swin Transformer identifying important areas affecting classification, with different attention focus across affected regions to facilitate model interpretability and diagnostic trust.

Moreover, Swin Transformer optimizes precision, FLOPs, and inference time, and it is an appropriate model for real-world agricultural usage where performance as well as efficiency is important. Its high generalization power and scalability further make it a strong deep learning model for challenging vision tasks.

To improve model interpretability, Grad-CAM visualization was used to examine the most important regions that affect classification decisions. Figure 8 shows heatmaps produced for correctly classified samples, indicating the areas of maximum activation. In the first image, the model has scattered attention, which may reflect sensitivity to background noise or irrelevant features. The second image shows a more focused attention, with activation centered around the diseased area, indicating better feature extraction. The third image again supports the model's credibility since it clearly shows the impacted regions, corresponding with expert markings. These images verify the model's capacity to distinguish between diseased and healthy areas, building trust in its output and offering guidance for further optimization.

V. Discussion

The classifying results obtained from the Swin Transformer model demonstrate outstanding performance on wheat diseases, with a 99.30% accuracy along with consistently high precision, recall, and F1 scores for all classes. Features in images captured through hierarchical structures of architectures and made use of shifted window attention mechanisms make the Swin Transformer efficient. This ability is critical in agricultural disease detection, where subtle visual differences can mean the difference between a correct diagnosis and an incorrect one. The robust precision and recall values of the model ensure that false positives and false negatives are minimized, which is important in agricultural applications to avoid the overuse of pesticides or the failure to detect diseases in time.

The model ensures that precision and recall are well balanced, meaning it would detect diseases with high sensitivity but at the same time avoid making a misclassification of healthy crops as diseased, which is a very important aspect of crop management; incorrect diagnosis can lead to unnecessary intervention, thereby wasting important resources or, worse still, undetected outbreaks of disease. The results also show that the Swin Transformer is robust enough to capture complex visual patterns that are present in agricultural disease datasets, where subtle and overlapping features are prevalent across different types of disease.

Furthermore, the robustness of the model against different conditions and the various data augmentation techniques applied, including rotation, flipping, and scaling, enhance the high generalization ability of the model. Moreover, the constant training and validation performance without significant overfitting can assure the model's reliability in real-world agricultural applications. Misclassifications in some classes between classes are visible in the confusion matrix, yet these can give important insights on how to further improve the performance of the model. Specifically, the confusion between Class 2 and Class 4 indicates that the model needs more training data or a stronger extraction of features to differentiate these diseases, which appear quite visually similar.

The Swin Transformer achieved impressive performance, but has notable limitations. It struggled to differentiate diseases with similar visual patterns, especially Class 2 and Class 4, which could be mitigated by diversifying the dataset or employing advanced feature extraction techniques. Further, robustness in real-world conditions is limited since many other factors such as lighting and weather conditions with a variety of symptoms will degrade accuracy, and it may require testing in more diverse scenarios or the inclusion of multi-modal data, such as weather and soil conditions. This model also requires interpretability for actual field applications; explainable AI techniques with XAI can be used to enhance transparency and thus instill confidence in users, like farmers, by explaining the reasons behind predictions.

A. Feasibility for Edge Deployment

Implementing deep learning models on resource-restricted edge devices, e.g., smartphones or UAV systems, involves significant computational efficiency concerns. Although the Swin Transformer attains excellent performance in wheat disease classification, its applicability for real-time crop monitoring via edge devices hinges on values such as FLOPs, inference time, and memory use.

Computational Complexity: Floating-point operations per second (FLOPs) give a model's computational load an estimate. Even as the Swin Transformer's shifted window attention mechanism is more efficient, its hierarchical architecture and having a large number of parameters result in greater FLOPs than lower-complexity models. This computational cost could be problematic for real-time deployment unless optimization is used.

Inference Time per Image: Real-time inference is important for agricultural use cases where timely feedback is required. Due to the large model size of the Swin Transformer, inference performance on low-power edge devices could be poor. Benchmarking inference latency on mobile processors or embedded AI accelerators is necessary to establish if the model is real-time capable.

Memory Usage: Memory limitations in edge devices can slow down deployment of big models such as the Swin Transformer. Its high memory requirements, needed to store parameters and operate on images, might surpass the limit of constrained devices, resulting in sluggish performance or system failure. Methods such as model quantization, pruning, and knowledge distillation may alleviate this concern by minimizing memory needs without substantially impacting accuracy.

B. Optimization Techniques for Edge Feasibility

To make Swin Transformer more edge-friendly, upcoming work will involve model compression methods, such as:

- *Quantization:* Precision reduction (e.g., from FP32 to INT8) to reduce computational and memory needs.
- *Pruning:* Elimination of redundant parameters to reduce model size and enhance efficiency.
- *Lightweight Alternatives:* Investigating transformer variants tailored to edge settings, e.g., MobileViT or EdgeFormer, to find an optimal balance between performance and efficiency.

All these optimizations will be crucial to making deep learning models deployable in practical agricultural use cases without going beyond hardware capabilities.

C. Comparison of Different Models

Our proposed Swin Transformer (Swin-Tiny) performs better than the conventional models such as ViT and DeiT for the case of wheat disease detection due to its capabilities to extract both local and global features through the shifted window self-attention mechanism. Although ViT and DeiT work well, they cannot utilize high-resolution agricultural images effectively at a reasonable computational cost.

In comparison with BEiT (BERT pre-trained Image Transformer), which uses masked image modeling, Swin Transformer continues to be faster and more performant in certain agricultural applications. ConvNeXt, a newer architecture as a direct ConvNet replacement for transformers, has offered competitive performance, but our Swin Transformer offers better performance for fine-grained disease classification, in which spatial information has to be processed more effectively.

Finally, EfficientNetv2 is efficient but lacks the flexibility of capturing longer-range dependencies that Swin Transformer has with its hierarchical and attention-based architecture. EfficientNetv2 may attain very high accuracy on generic image classification tasks, but there is improved scalability and fine-tuning capability for domain-specific tasks such as wheat disease detection through Swin Transformer's architecture.

Although EfficientNetv2 and ConvNeXt are more computationally efficient, Swin Transformer provides superior flexibility in handling complex image data, offering the best of both worlds in terms of accuracy, generalization, and scalability in agricultural applications. Our work thus builds upon and differs from these existing approaches by introducing the shifted window attention mechanism, which excels in localized attention and hierarchical feature extraction, making it particularly suited for detecting subtle changes in plant disease symptoms.

D. Limitations

Although the Swin Transformer model attained high accuracy, it has a number of challenges that need to be overcome in order to deploy it in real-world scenarios. Its most significant limitation is its computational complexity, which limits its use in mobile and IoT-based farming systems without optimization. Its deployment on the edge devices involves efficient architectures with regard to performance vs. power consumption. Also, environmental variation presents challenges since the accuracy of the model reduces under low light conditions, unconventional camera angles, and changing weather conditions. Providing robustness across varied real-world situations is a necessity for reliable disease detection. Moreover, scalability of the model is still constrained, as it is mostly trained for wheat disease classification. Enlarging its capacity to recognize several crop diseases through self-supervised learning could enhance generalization across wider agricultural applications. Overcoming these limitations will be essential to making the Swin Transformer a viable solution for precision farming, allowing for scalable, real-time, and effective disease monitoring.

VI. CONCLUSION & FUTURE WORK

This study demonstrates the efficacy of the Swin Transformer model in classifying wheat diseases, achieving a high F1-score of 99.30%. The model's hierarchical design and shifted window attention enable the identification of subtle visual patterns, making it highly suitable for agricultural applications. By effectively modeling both local and global contextual information, the Swin Transformer exhibits strong performance across various wheat disease classes. Its high precision and recall scores indicate reduced false positives and false negatives–critical factors for accurate disease diagnosis and management in real-world farming scenarios. Additionally, the model's computational efficiency and generalization capability across diverse datasets position it as a strong candidate for real-time deployment in precision agriculture. The proposed model can be integrated into practical solutions, such as mobile-based diagnostic tools or UAV-based monitoring systems. Such integration has the potential to improve early disease intervention, support crop health monitoring, and ultimately enhance yield while minimizing losses due to disease outbreaks.

Future work will focus on extending the application of the Swin Transformer to the detection of other plant diseases and its deployment within UAV-based intelligent farming systems. We plan to enhance the model's robustness and generalizability by expanding the dataset to include multi-location wheat disease samples under varying environmental conditions. Additionally, integration with autonomous platforms such as agricultural robots and drones will enable large-scale, real-time disease monitoring, reducing manual labor and facilitating precise interventions.

To increase model interpretability, future efforts will explore SHAP-based feature attribution techniques, which will provide deeper insights into the model's decision-making process. Enhancing interpretability will not only improve transparency but also build trust in AI-based disease diagnostics, aiding broader adoption in agricultural practice. These advancements will contribute to scalable, efficient, and intelligent solutions in sustainable agriculture and strengthen the future of automated farming systems.

We also intend to evaluate the Swin Transformer on wheat datasets collected from diverse geographical regions and climatic zones to improve its robustness and adaptability. Further investigation into the model's performance across different wheat species and agricultural practices will enhance its practical utility, supporting comprehensive disease detection in varied real-world conditions.

ACKNOWLEDGMENTS

The open access funding is provided by Qatar National Library.

REFERENCES

[1] K. Pawlak and M. Kołodziejczak, "The role of agriculture in ensuring food security in developing countries: Considerations in the context of the problem of sustainable food production," *Sustainability*, vol. 12, no. 13, p. 5488, 2020.

[2] M. Zhuang et al., "The sustainability of staple crops in China can be substantially improved through localized strategies," *Renew. Sustain. Energy Rev.*, vol. 154, Feb. 2022, Art. no. 111893.

[3] M. Kayim, H. Nawaz, and A. Alsalmo, "Fungal diseases of wheat," in *Wheat-Recent Advances*. London, U.K.: IntechOpen, 2022.

[4] O. Erenstein, M. Jaleta, K. A. Mottaleb, K. Sonder, J. Donovan, and H.-J. Braun, "Global trends in wheat production, consumption and trade," in *Wheat Improvement: Food security in a Changing Climate*. Cham, Switzerland: Springer Int., 2022, pp. 47–66.

[5] R. P. Singh, H. M. William, J. Huerta-Espino, and G. Rosewarne, "Wheat rust in Asia: Meeting the challenges with old and new technologies," in *Proc. 4th Int. Crop Sci. Congr.*, vol. 26, Sep. 2004, pp. 1–13.

[6] S. N. Wegulo, P. S. Baenziger, J. H. Nopsa, W. W. Bockus, and H. Hallen-Adams, "Management of fusarium head blight of wheat and barley," *Crop Protect.*, vol. 73, pp. 100–107, Jul. 2015.

[7] H. Wieser, P. Koehler, and K. A. Scherf, "The two faces of wheat," *Frontiers in Nutrition*, vol. 7, Oct. 2020, Art. no. 517313.

[8] G. W. Norton, J. Alwang, and W. A. Masters, *Economics of Agricultural Development: World Food Systems and Resource Use*. Abingdon, U.K.: Routledge, 2021.

[9] M. S. Manavadaria, D. Sasikala, R. Rajamohanan, V. M. Agone, and G. C. Babu, "Automated detection and alerting system for wheat leaf diseases using the VGG16 deep learning model," in *Proc. Int. Conf. Data Sci. Netw. Security (ICDSNS)*, Jul. 2024, pp. 1–5.

[10] D. Suri, S. Saksenaa, U. Sehgal, and R. Garg, "Disease classification in wheat from images using cnn," in *Proc. 13th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, 2023, pp. 566–571.

[11] N. Dahiya, G. Singh, S. Singh, and V. Sood, "Crop land assessment with deep neural network using hyperspectral satellite dataset," in *Hyperautomation in Precision Agriculture*. Amsterdam, The Netherlands: Elsevier, 2025, pp. 159–167.

[12] S. Nigam et al., "Wheat disease severity estimation: A deep learning approach," in *Proc. Int. Conf. Internet Things Connected Technol.*, 2021, pp. 185–193.

[13] Y. Borhani, J. Khoramdel, and E. Najafi, "A deep learning based approach for automated plant disease classification using vision transformer," *Sci. Rep.*, vol. 12, no. 1, 2022, Art. no. 11554.

[14] C. Gu, D. Wang, H. Zhang, J. Zhang, D. Zhang, and D. Liang, "Fusion of deep convolution and shallow features to recognize the severity of wheat fusarium head blight," *Front. Plant Sci.*, vol. 11, Jan. 2021, Art. no. 599886.

[15] S. Mehta, V. Kukreja, and S. Vats, "Advancing agricultural practices: Federated learning-based CNN for mango leaf disease detection," in *Proc. 3rd Int. Conf. Intell. Technol. (CONIT)*, 2023, pp. 1–6.

[16] H. Zhang and G. Ren, "Intelligent leaf disease diagnosis: Image algorithms using swin transformer and federated learning," *Visual Comput.*, vol. 41, pp. 4815–4838, May 2025.

[17] R. Bandi, S. Swamy, and C. Arvind, "Leaf disease severity classification with explainable artificial intelligence using transformer networks," *Int. J. Adv. Technol. Eng. Exploration*, vol. 10, no. 100, p. 278, 2023.

[18] K. Rethik and D. Singh, "Attention based mapping for plants leaf to classify diseases using vision transformer," in *Proc. 4th Int. Conf. Emerg. Technol. (INCET)*, 2023, pp. 1–5.

[19] V. Singhi, D. Kumar, and V. Kukreja, "Integrated YOLOv4 deep learning pretrained model for accurate estimation of wheat rust disease severity," in *Proc. 2nd Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC)*, 2023, pp. 489–494.

[20] A. Chougui, A. Moussaoui, and A. Moussaoui, "Plant-leaf diseases classification using CNN, CBAM and vision transformer," in *Proc. 5th Int. Symp. Inform. Appl. (ISIA)*, 2022, pp. 1–6.

[21] I. Kunduracioglu and I. Pacal, "Advancements in deep learning for accurate classification of grape leaves and diagnosis of grape diseases," *J. Plant Dis. Protect.*, vol. 131, no. 3, pp. 1061–1080, 2024.

[22] J. Leukel, T. Zimpel, and C. Stumpe, "Machine learning technology for early prediction of grain yield at the field scale: A systematic," *Comput. Electron. Agr.*, vol. 207, Apr. 2023, Art. no. 107721.

[23] I. Pacal and G. Işık, "Utilizing convolutional neural networks and vision transformers for precise corn leaf disease identification," *Neural Comput. Appl.*, vol. 37, no. 4, pp. 2479–2496, Feb. 2025.

[24] P. S. Thakur, P. Khanna, T. Sheorey, and A. Ojha, "Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT," 2022, *arXiv:2207.07919*.

[25] C. Zhou, Y. Zhong, S. Zhou, J. Song, and W. Xiang, "Rice leaf disease identification by residual-distilled transformer," *Eng. Appl. Artif. Intell.*, vol. 121, May 2023, Art. no. 106020.

[26] J. Su et al., "Aerial visual perception in smart farming: Field study of wheat yellow rust monitoring," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2242–2249, Mar. 2021.

[27] H.-T. Thai, K.-H. Le, and N. L.-T. Nguyen, "FormerLeaf: An efficient vision transformer for cassava leaf disease detection," *Comput. Electron. Agr.*, vol. 204, Jan. 2023, Art. no. 107518.

[28] C. Nguyen, V. Sagan, J. Skobalski, and J. I. Severo, "Early detection of wheat yellow rust disease and its impact on terminal yield with multi-spectral UAV-imagery," *Remote Sens.*, vol. 15, no. 13, p. 3301, 2023.

[29] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[30] B. Espejo-Garcia, H. Panoutsopoulos, E. Anastasiou, F. J. Rodríguez-Rigueiro, and S. Fountas, "Top-tuning on transformers and data augmentation transferring for boosting the performance of weed identification," *Comput. Electron. Agr.*, vol. 211, Art. no. 108055, Aug. 2023.

[31] S. M. Javidan, A. Banakar, K. A. Vakilian, and Y. Ampatzidis, "Diagnosis of grape leaf diseases using automatic k-means clustering and machine learning," *Smart Agr. Technol.*, vol. 3, Feb. 2023, Art. no. 100081.

[32] D. Das, M. Singh, S. S. Mohanty, and S. Chakravarty, "Leaf disease detection using support vector machine," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, 2020, pp. 1036–1040.

[33] W. Wei, Y. Tian-Le, and L. Rui, "Detection and enumeration of wheat grains based on a deep learning method under various scenarios and scales," *J. Integrative Agr.*, vol. 19, no. 8, pp. 1998–2008, 2020.

[34] A. Bhatia, A. Chug, and A. P. Singh, "Plant disease detection for high dimensional imbalanced dataset using an enhanced decision tree approach," *Int. J. Future Gener. Commun. Netw.*, vol. 13, no. 4, pp. 71–78, 2020.

[35] S. K. Sahu and M. Pandey, "An optimal hybrid multiclass SVM for plant leaf disease detection using spatial fuzzy c-means model," *Expert Syst. Appl.*, vol. 214, Mar. 2023, Art. no. 118989.

[36] M. Govardhan and M. Veena, "Diagnosis of tomato plant diseases using random forest," in *Proc. Global Conf. Adv. Technol. (GCAT)*, 2019, pp. 1–5.

[37] H.-T. Thai, N.-Y. Tran-Van, and K.-H. Le, "Artificial cognition for early leaf disease detection using vision transformers," in *Proc. Int. Conf. Adv. Technol. Commun. (ATC)*, 2021, pp. 33–38.

[38] M. A. Islam, M. N. R. Shuvo, M. Shamsojjaman, S. Hasan, M. S. Hossain, and T. Khatun, "An automated convolutional neural network based approach for paddy leaf disease detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 1, pp. 1–9, 2021.

[39] Z. Jiang, Z. Dong, W. Jiang, and Y. Yang, "Recognition of rice leaf diseases and wheat leaf diseases based on multi-task deep transfer learning," *Comput. Electron. Agr.*, vol. 186, Jul. 2021, Art. no. 106184.

[40] S. S. Chouhan, U. P. Singh, and S. Jain, "Automated plant leaf disease detection and classification using fuzzy based function network," *Wireless Personal Commun.*, vol. 121, no. 3, pp. 1757–1779, 2021.

[41] R. Deshpande and H. Patidar, "Tomato plant leaf disease detection using generative adversarial network and deep convolutional neural network," *Imag. Sci. J.*, vol. 70, no. 1, pp. 1–9, 2022.

[42] P. V. Yeswanth, S. Deivalakshmi, S. George, and S.-B. Ko, "Residual skip network-based super-resolution for leaf disease detection of grape plant," *Circuits, Syst. Signal Process.*, vol. 42, no. 11, pp. 6871–6899, 2023.

[43] V. P. Gaikwad and V. Musande, "Wheat disease detection using image processing," in *Proc. 1st Int. Conf. Intell. Syst. Inf. Manage. (ICISIM)*, 2017, pp. 110–112.

[44] Z. Mi, X. Zhang, J. Su, D. Han, and B. Su, "Wheat stripe rust grading by deep learning with attention mechanism and images from mobile devices," *Front. Plant Sci.*, vol. 11, Sep. 2020, Art. no. 558126.

[45] N. Mishra, J. Seetha, A. G. D. Kumar, M. S. Menon, and A. Ravuri, "Design an ant lion-based YOLO-v5 model for prediction and classification of paddy leaf disease," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 6, pp. 599–612, 2023.

[46] H. Alshammari, K. Gasmi, I. Ben Ltaifa, M. Krichen, L. Ben Ammar, and M. A. Mahmood, "Olive disease classification based on vision transformer and CNN models," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, 2022, Art. no. 3998193.

[47] S. Hossain, M. T. Reza, A. Chakrabarty, and Y. J. Jung, "Aggregating different scales of attention on feature variants for tomato leaf disease diagnosis from image data: A transformer driven study," *Sensors*, vol. 23, no. 7, p. 3751, 2023.

[48] A. Smetanin, A. Uzhinskiy, G. Ososkov, P. Goncharov, and A. Nechaevskiy, "Deep learning methods for the plant disease detection platform," in *Proc. AIP Conf.*, 2021, Art. no. 60006.

[49] L. Xu et al., "Wheat leaf disease identification based on deep learning algorithms," *Physiol. Mol. Plant Pathol.*, vol. 123, Jan. 2023, Art. no. 101940.

[50] Q. Pan, M. Gao, P. Wu, J. Yan, and S. Li, "A deep-learning-based approach for wheat yellow rust disease recognition from unmanned aerial vehicle images," *Sensors*, vol. 21, no. 19, pp. 1–17, 2021.

[51] R. Sujatha, J. M. Chatterjee, N. Jhanjhi, and S. N. Brohi, "Performance of deep learning vs machine learning in plant leaf disease detection," *Microprocess. Microsyst.*, vol. 80, Feb. 2021, Art. no. 103615.

[52] Y. Li et al., "Semantic segmentation of wheat stripe rust images using deep learning," *Agronomy*, vol. 12, no. 12, p. 2933, 2022.

[53] V. Kukreja and D. Kumar, "Automatic classification of wheat rust diseases using deep convolutional neural networks," in *Proc. 9th Int. Conf. Rel. Infocom Technol. Optim. (Trends and Future Directions) (ICRITO)*, 2021, pp. 1–6. é

[54] Y. Gao, H. Wang, M. Li, and W.-H. Su, "Automatic tandem dual blendmask networks for severity assessment of wheat fusarium head blight," *Agriculture*, vol. 12, no. 9, p. 1493, 2022.

[55] E. Ennadifi, S. Laraba, D. Vincke, B. Mercatoris, and B. Gosselin, "Wheat diseases classification and localization using convolutional neural networks and GradCAM visualization," in *Proc. Int. Conf. Intell. Syst. Comput. Vis. (ISCV)*, 2020, pp. 1–5.

[56] T. Hayit, H. Erbay, F. Varçın, F. Hayit, and N. Akci, "Determination of the severity level of yellow rust disease in wheat by using convolutional neural networks," *J. Plant Pathol.*, vol. 103, no. 3, pp. 923–934, 2021.

[57] R. R. Patil, S. Kumar, S. Chiwhane, R. Rani, and S. K. Pippal, "An artificial-intelligence-based novel rice grade model for severity estimation of rice diseases," *Agriculture*, vol. 13, no. 1, pp. 1–19, 2023.

[58] U. Shafi et al., "Embedded ai for wheat yellow rust infection type classification," *IEEE Access*, vol. 11, pp. 23726–23738, 2023.

[59] S. Jain et al., "Automatic rice disease detection and assistance framework using deep learning and a chatbot," *Electronics*, vol. 11, no. 14, pp. 1–18, 2022.

[60] M. A. Genaev, E. S. Skolotneva, E. I. Gultyaeva, E. A. Orlova, N. P. Bechtold, and D. A. Afonnikov, "Image-based wheat fungi diseases identification by deep learning," *Plants*, vol. 10, no. 8, p. 1500, 2021.

[61] R. Mumtaz, M. H. Maqsood, I. U. Haq, U. Shafi, Z. Mahmood, and M. Mumtaz, "Integrated digital image processing techniques and deep learning approaches for wheat stripe rust disease detection and grading," *Decis. Anal. J.*, vol. 8, Sep. 2023, Art. no. 100305.

[62] B. S. Bari et al., "A real-time approach of diagnosing rice leaf disease using deep learning-based faster RCNN framework," *PeerJ Comput. Sci.*, vol. 7, pp. 1–27, Apr. 2021.

[63] L. Goyal, C. M. Sharma, A. Singh, and P. K. Singh, "Leaf and spike wheat disease detection & classification using an improved deep convolutional architecture," *Inform. Med. Unlocked*, vol. 25, 2021, Art. no. 100642.

[64] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Front. Plant Sci.*, vol. 7, p. 1419, Sep. 2016. [Online]. Available: https://github.com/spMohanty/PlantVillage-Dataset

[65] A. K. Singh and A. Misra. "Wheat leaf dataset." 2020. Accessed: Apr. 25, 2025. [Online]. Available: https://www.kaggle.com/datasets/arindam235/wheat-leaf-disease-dataset

[66] H. Getachew, "Wheat leaf dataset: Mendeley Data, Version 1," 2021. Accessed: Apr. 25, 2025. [Online]. Available: https://data.mendeley.com/datasets/wgd66f8n6h/1

[67] W. Bao, J. Zhao, G. Hu, D. Zhang, L. Huang, and D. Liang, "Identification of wheat leaf diseases and their severity based on elliptical-maximum margin criterion metric learning," *Sustain. Comput. Inform. Syst.*, vol. 30, Jun. 2021, Art. no. 100526.

[68] X. Fang, T. Zhen, and Z. Li, "Lightweight multiscale CNN model for wheat disease detection," *Appl. Sci.*, vol. 13, no. 9, p. 5801, 2023.

[69] M. Dong, S. Mu, A. Shi, W. Mu, and W. Sun, "Novel method for identifying wheat leaf disease images based on differential amplification convolutional neural network," *Int. J. Agr. Biol. Eng.*, vol. 13, no. 4, pp. 205–210, 2020.

[70] W. Haider, A.-U. Rehman, N. M. Durrani, and S. U. Rehman, "A generic approach for wheat disease classification and verification using expert opinion for knowledge-based decisions," *IEEE Access*, vol. 9, pp. 31104–31129, 2021.

Muhammad Khubaib is currently pursuing the M.S. degree in computer science with The University of Lahore, Pakistan. His research focuses on advanced applications of deep learning and image processing in healthcare, including AI-based disease detection and computer vision solutions. His work encompasses innovative approaches to healthcare challenges, such as developing deep learning models for detecting diseases and improving diagnostic accuracy. With publications in recognized journals, he is dedicated to leveraging artificial intelligence for meaningful advancements in technology and healthcare.

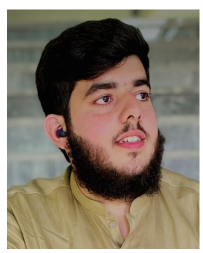
Tanzila Kehkashan received the master's degree from the University of Central Punjab, Pakistan, and the Ph.D. degree from Universiti Teknologi Malaysia, Malaysia. She is an Assistant Professor with the Faculty of Information Technology, The University of Lahore, Pakistan. In addition to her teaching role, she serves as the Principal Investigator with the Vision and Language Computing Matrix Lab, The University of Lahore, where she leads research projects in the areas of artificial intelligence, computer vision, and natural language processing. Her research interests include artificial intelligence, machine learning, deep learning, computer vision, natural language processing, and the development of innovative solutions for real-world problems in healthcare, agriculture, education, and cybersecurity. Her dedication to both research and academic practice reflects her passion for the field of computer science. Her work has been published in prestigious journals and conference proceedings, showcasing her commitment to advancing research in these fields.

Maha Abdelhaq (Member, IEEE) received the B.Sc. degree in computer science and the M.Sc. degree in securing wireless communications from the University of Jordan, Jordan, in 2006 and 2009, respectively, and the Ph.D. degree from the Faculty of Information Science and Technology, National University of Malaysia, Malaysia, in 2014. She is currently an Associate Professor with the College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Saudi Arabia. Her research interests include vehicular networks, MANET routing protocols, artificial immune systems, network security, and intelligent computational. She is a member of ACM and the International Association of Engineers.

Muhammad Asghar Khan (Senior Member, IEEE) received the B.E. degree in electronic engineering from IQRA University, Pakistan, the M.S. degree in electrical engineering from UET Taxila, Pakistan, and the Ph.D. degree in electronic engineering from the School of Engineering and Applied Sciences, ISRA University, Pakistan. He is an Associate Research Professor in the Department of Electrical Engineering with Prince Mohammad Bin Fahd University (PMU), Saudi Arabia. Prior to joining PMU, he served as an Associate Professor and the Head of the Electrical Engineering Department, Hamdard University, Islamabad. During his tenure, he spearheaded the successful accreditation of the Electrical Engineering Program under the Washington Accord for the 2019–2022 intake batches. His in-depth expertise in the accreditation criteria of the Washington Accord and ABET has significantly enhanced academic and institutional standards. He also served as a Researcher with the Smart Systems Engineering Lab, Prince Sultan University, Saudi Arabia, from June 2022 to June 2023. He is a distinguished researcher, having authored and co-authored over 100 technical and review articles published in leading journals. His research interests are focused on the development and integration of advanced technologies in drones/UAVs. His work encompasses network architectures, security frameworks, platform innovations, and cutting-edge applications for UAV systems. His commitment to academic excellence and impactful research demonstrates his dedication to advancing technology and shaping the future of modern engineering solutions.

Muhammad Zaman (Member, IEEE) is currently pursuing the Master of Science degree in computer science with the COMSATS University of Islamabad. He is a Committed Lecturer with the Superior University. He, driven by an unwavering dedication to the progression of knowledge, possesses extensive research acumen in numerous fields, including but not limited to medical image processing, remote sensing, and natural language processing. The individual's substantial research pursuits and contributions demonstrate a profound enthusiasm for artificial intelligence, machine learning, deep learning, computer vision, and reinforcement learning. His numerous publications in computer vision have substantiated his contributions to the field. Furthermore, he has served as a mentor and guide to a considerable number of M.S. students as they have completed their research theses and papers. This demonstrates his dedication to cultivating the subsequent generation of scholars and innovators. He is Distinguished Member of IEEE.

Imran Ashraf is currently pursuing the bachelor's degree in software engineering with The University of Lahore, Pakistan. As a Research Member of the Vision and Language Computing Matrix Lab, The University of Lahore, he supports innovative projects in artificial intelligence and machine learning under expert supervision. His contributions focus on computational systems and data-driven algorithms, aiding the development of solutions for real-world applications. He leverages his software engineering expertise to enhance collaborative research efforts, ensuring robust project outcomes. His academic interests include advanced AI technologies and their impact across diverse domains. Committed to academic excellence, he has engaged in projects presented within research forums, strengthening his expertise in AI-driven solutions. His dedication to supporting cutting-edge research reflects his drive to contribute to technological advancements through teamwork and continued learning.



Abdul Rehman is currently pursuing the undergraduate degree in computer science from The University of Lahore, Pakistan, with a keen interest in machine learning, natural language processing, and medical imaging. His research area is the creation of AI-based solutions for healthcare, such as diabetic foot ulcer detection, in which he has experimented with SSD and YOLO architectures to improve detection accuracy. Outside of medical diagnostics, he is highly interested in sign language detection and classification of brain tumors, using deep learning to improve accessibility and medical imaging analysis. He is passionate about innovation and seeks to use AI and computational intelligence to drive revolutionary breakthroughs in healthcare and assistive technologies.

Adnan Akhunzada (Senior Member, IEEE) brings 15 years of expertise in Research and Development at the nexus of the ICT industry and academia. Renowned for his high-impact publications, U.S. patents, and commercial products, his patented innovations in cybersecurity and AI have secured multi-million-dollar projects with global entities, such as Vinnova and EU Horizon. In 2023, Stanford University recognized him as one of the top 2% scientists globally for his outstanding scholarly contributions. Leveraging his robust cybersecurity skills and advanced technological knowledge, he excels in solving industrial challenges and developing state-of-the-art security tools and frameworks. His expertise spans cybersecurity and AI, secure future Internet, secure and dependable software defined networks, and large-scale distributed systems (including Cloud, Fog, Edge, IoT, IoE, IIoT, and CPS). Additionally, his work on lightweight cryptographic communication protocols, QoS/QoE, and adversarial machine learning is shaping the future of secure and dependable systems. He is a Distinguished Senior Member of IEEE and a Professional Member of ACM.