Transformer-based Models for foliar Maize Diseases Classification in Real Field Conditions: Case Study of Kandahar, Afghanistan

R. Rafi[1], A. Soleimanipour[2*], A. Rezaei Asl[2]

1- Agricultural Engineering Department, Plant Science Faculty, Afghanistan National Agricultural Science and Technology University (ANASTU), Kandahar, Afghanistan
2- Department of Biosystem Engineering, Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran
(*- Corresponding Author Email: asoleimani@gau.ac.ir)
https://doi.org/10.22067/jam.2026.97474.1465

Abstract

Accurate and timely detection of maize (*Zea mays* L.) foliar diseases is essential for improving crop productivity and ensuring food security. This study evaluates the performance of state-of-the-art deep learning models for automated maize leaf disease classification using a real-field dataset collected from agricultural farms in Kandahar Province, Afghanistan. The dataset was acquired under natural field conditions, preserving complex backgrounds, variable illumination, and diverse leaf orientations to reflect practical farming environments. Multiple advanced transformer-based deep learning architectures, including Vision Transformer (ViT), and hybrid Multi-axis Vision Transformer (MaxViT-tiny, MaxViT-small, and MaxViT-base) were investigated and compared with EfficientNet (B5, B6, and B7) models. Experimental results demonstrate that modern CNN-based and transformer-based architectures achieve up to 97% accuracy, despite challenging real-world conditions. Among the evaluated models, EfficientNet-B7, ViT, and MaxViT-Tiny attained the highest accuracy of 97%, with MaxViT-Tiny recording the highest precision (0.9741), indicating exceptional class discrimination, particularly for challenging Spot and Blight classes under real-field variability. Minor misclassifications were primarily observed between visually similar disease classes, while overall precision and recall remained consistently high. The findings confirm the effectiveness of hybrid and attention-based architectures for real-world maize disease detection and demonstrate their feasibility for real-world agricultural applications. By leveraging authentic field data, this study provides practical and ecologically valid insights that support early disease diagnosis, reduced manual inspection, and improved sustainable crop management.

Keywords: EfficientNet, Field dataset, leaf disease, Maize, MaxViT, ViT

Introduction

Agriculture remains a cornerstone of global development, underpinning food security, employment, and economic stability amid a rapidly growing world population (Eli-Chukwu, 2019; Koroshi Talab, Mohammad Zamani, & Gholami Par-Shokohi, 2025). However, conventional farming practices—characterised by manual crop inspection and broad-spectrum pesticide application—are increasingly recognised as inefficient and environmentally unsustainable, limiting their capacity to meet rising production demands (Keerthivasan *et al.*, 2026; Shamloo, Soleimanipour, & Rezaei Asl, 2025). Among staple crops, maize (*Zea mays* L.) occupies a strategic position in global agriculture, functioning not only as a primary food source but also as livestock feed and industrial raw material. Its economic and nutritional importance makes safeguarding maize productivity a priority for national and global food systems (Entuni & Zulcaffle, 2022; Idress, Gadalla, Öztekin, & Baitu, 2024).

Maize production is particularly vulnerable to foliar diseases such as grey leaf spot, common rust, and northern leaf blight, which can substantially reduce yield and market quality if not identified at early stages (Bruns, 2017; Ma *et al.*, 2022). Empirical field investigations and epidemiological analyses consistently demonstrate that delayed diagnosis significantly amplifies both economic losses and disease spread (Panigrahi, Das, Sahoo, & Moharana, 2020; Waheed *et al.*, 2020). Consequently, timely and precise detection has become a fundamental requirement for sustainable maize cultivation (Pacal & Işık, 2025). Traditional diagnostic approaches—primarily visual inspection by trained experts followed by laboratory validation—remain widely practised but present critical limitations. These methods are labour-intensive, time-consuming, and inherently subjective, leading to variability in diagnostic accuracy (Fraiwan, Faouri, & Khasawneh, 2022). Moreover, they are impractical for large-scale monitoring across geographically dispersed farms (Setiawan, Rochman, Satoto, & Rachmad, 2022).

To address these limitations, artificial intelligence (AI), particularly deep learning (DL), has emerged as a transformative tool for automated plant disease detection. Convolutional neural networks (CNNs) have demonstrated strong capability in extracting discriminative local features from leaf images, enabling accurate classification across multiple crop species (Dhaka *et al.*, 2021). Subsequent advancements have incorporated optimised training strategies and lightweight architectures to improve scalability and deployment feasibility (El Sakka, Mothe, & Ivanovici, 2024). More recently, multi-task deep learning frameworks have extended disease recognition systems by integrating simultaneous classification and severity estimation, thereby reducing reliance on expert intervention and supporting precision agriculture workflows (Dai Xia, Zhu, & Che, 2023; Rashid, Aslam, Aziz, & Aldehim, 2024).

Despite these advances, CNN-based models inherently emphasise localised receptive fields and may struggle to capture long-range spatial dependencies characteristic of complex disease patterns (Alshammari *et al.*, 2022). Vision Transformers (ViTs) were introduced to overcome this limitation through self-attention mechanisms that model global contextual relationships within images (Ullah *et al.*, 2024). Hybrid architectures, including the Multi-axis Vision Transformer (MaxViT), further combine convolutional inductive biases with transformer-based global attention, offering improved representational capacity and robustness under heterogeneous imaging conditions (Nishankar *et al.*, 2025; Pranta *et al.*, 2025).

Notwithstanding the methodological progress, important research gaps persist. A substantial portion of existing studies rely on controlled or background-cleaned datasets, most notably *PlantVillage*, which do not adequately represent real-field complexity. Such datasets typically minimise environmental variability, thereby limiting the ecological validity and generalisation capacity of trained models when deployed in practical agricultural settings. Field environments introduce challenges, including non-uniform illumination, cluttered natural backgrounds, leaf occlusion, scale variation, and arbitrary orientation, which are the factors that systematically degrade model performance. Although CNNs dominate the literature, systematic comparative evaluations of CNN-based, transformer-based, and hybrid architectures on authentic, small-scale field datasets from underrepresented regions remain limited.

This study addresses these gaps by comprehensively evaluating state-of-the-art architectures (including EfficientNet variants, ViT, and MaxViT) using a real-field maize leaf dataset collected directly from farms in Kandahar Province, Afghanistan. The primary objectives are to: (1) compare model performance under authentic field conditions, (2) analyse the influence of architectural and training configurations on classification metrics, and (3) identify computationally efficient models suitable for deployment in resource-constrained precision agriculture systems.

The remainder of this paper is organised as follows: Section 2 reviews related work on deep learning-based maize disease detection and highlights persistent gaps; Section 3 describes the field-collected dataset and experimental methodology; Section 4 presents quantitative results, learning curves, confusion matrices, and in-depth discussion; and Section 5 discusses limitations and future research directions.

Related Works

Recent advances in deep learning-based models have improved crop disease detection performance significantly (Ahmadi, 2025). Rajeena, Su, Moustafa, and Ali (2023) developed an EfficientNet-based classification framework incorporating image preprocessing and variable-tuned model training, reaching a recognition accuracy of 98.85%, significantly outperforming earlier approaches. Similarly, Masood *et al.* (2023) proposed MaizeNet, a hybrid architecture combining Faster R-CNN with a ResNet-50 backbone and spatial-channel attention. Their model achieved 97.89% accuracy and a mAP of 0.94 on the Corn Disease and Severity dataset. In response to the challenge of field variability, Yang, Yao, and Teng (2024) enhanced the YOLOv8 model by incorporating Slim-neck and GAM attention modules. Their modified model attained 95.18% precision, 89.11% recall, and an mAP50 of 94.65%, demonstrating superior accuracy and robustness in maize leaf spot disease identification.

Transformer-based approaches are also emerging in plant disease detection. Pranta *et al.* (2025) introduced MaxViT-XSLD, a hybrid Vision Transformer model integrating multi-axis attention and MBConv layers for the joint classification of soybean seed and leaf diseases. Achieving over 99% accuracy on augmented datasets, the model also provided explainable predictions through XAI techniques and was deployed in the SoyScan real-time diagnostic tool.

ViT-RoT was proposed by Nishankar *et al.* (2025) as a benchmarking framework for evaluating various ViT architectures. This study identified ConvNeXt-Small and Swin-Small as the top-performing models across standard benchmarks, offering insights into ViT selection for agricultural diagnostics. To enable deployment on resource-constrained devices, Ullah *et al.* (2024) developed AppViT, a hybrid model that combines convolutional layers with ViT blocks. AppViT achieved 96.38% precision, 95.9% recall, and a 96.3% F1-score on the Plant Pathology 2021-FGVC8 dataset, outperforming ResNet-50 and EfficientNet-B3, while maintaining a lightweight structure with just 1.3 million parameters. Further studies by Entuni and Zulcaffle (2022) demonstrated classification accuracies of up to 95.11% using DenseNet-201 and ResNet-50 on corn leaf disease datasets. Amin, Darwish, Hassanien, and Soliman (2022) achieved 98.56% accuracy by combining EfficientNetB0 and DenseNet121, illustrating the effectiveness of transfer learning and feature fusion in plant disease classification.

Although these studies report high classification accuracies, most rely on controlled or benchmark datasets and do not evaluate model robustness under real-field conditions. Furthermore, few studies provide a systematic comparison between CNN-based, transformer-based, and hybrid architectures using authentic field data, which limits their practical generalisability. This study addresses these gaps by conducting a comparative evaluation under realistic agricultural conditions.

Materials and Methods

Experimental Environment

This research explores the automated detection and classification of maize leaf diseases utilising deep learning models. Model development and evaluation were conducted using the PyTorch framework. The methodological framework includes dataset preparation, image preprocessing, architectural design of the models, training procedures, and the application of comprehensive evaluation metrics. All experiments were performed using Google Colab Pro, utilising an NVIDIA Tesla T4 GPU with 16 GB VRAM and 12 GB RAM to ensure computational efficiency during model training and evaluation. The implementation was conducted in Python 3.10 with key deep learning frameworks, including PyTorch 2.0.1, the timm library for Vision Transformer models, and scikit-learn for evaluation metrics. Training and evaluation workflows were executed in a single GPU-accelerated Colab notebook. This configuration aligns with recent research efforts that have employed Google Colaboratory in conjunction with Tesla T4 GPUs to facilitate the training of deep learning models for plant disease detection and classification (Jain *et al.*, 2022; Jaiswal, Nagwanshi, Kumar, Gaurav, & Watti, 2024).

Dataset Description

The maize leaf images were collected from multiple fields across Kandahar Province, Afghanistan, under typical local environmental conditions. The region experiences a semi-arid climate, with average daytime temperatures ranging from 15°C to 38°C during the growing season, relative humidity between 20%–45%, and sporadic rainfall. These conditions introduce natural variations in lighting, leaf moisture, and background soil, providing realistic field scenarios for disease classification.

The datasets employed in this study were carefully curated to ensure both scientific rigour and practical relevance in maize leaf disease classification. The complementary sources of data were used: a field dataset collected from Kandahar Province, Afghanistan, which introduces real-world environmental variability. Image acquisition was conducted during July and August 2025, covering different times of day (morning and afternoon) to capture lighting variations and seasonal conditions. Data were collected across four representative agricultural districts (including Dand, Panjwayi, Arghandab, and Zheray) using a Samsung SM-A715F smartphone equipped with a 12-megapixel camera (resolution: 3472 × 3472 pixels). Each image was stored in JPEG format with embedded geolocation metadata (latitude and longitude), enabling precise mapping of sampling locations. The field dataset was divided into three primary classes: Healthy maize leaves, leaves infected with Spot disease, and leaves infected with Northern Leaf Blight. The workflow of data collection includes preprocessing and organising, as shown in Figure 1. Previous studies have shown that common rust development requires warm temperatures, high relative humidity, and sufficient leaf wetness to facilitate spore germination and infection (Chang, Wei, Liu, Zhao, & Shi, 2025; Sopialena, Suyadi, & Noor, 2022). However, the agro-climatic conditions of Kandahar Province which characterised by low rainfall, high temperatures, and minimal humidity during the cropping season, are not conducive to the establishment and spread of rust infections; therefore, rust disease was not included in this field dataset.

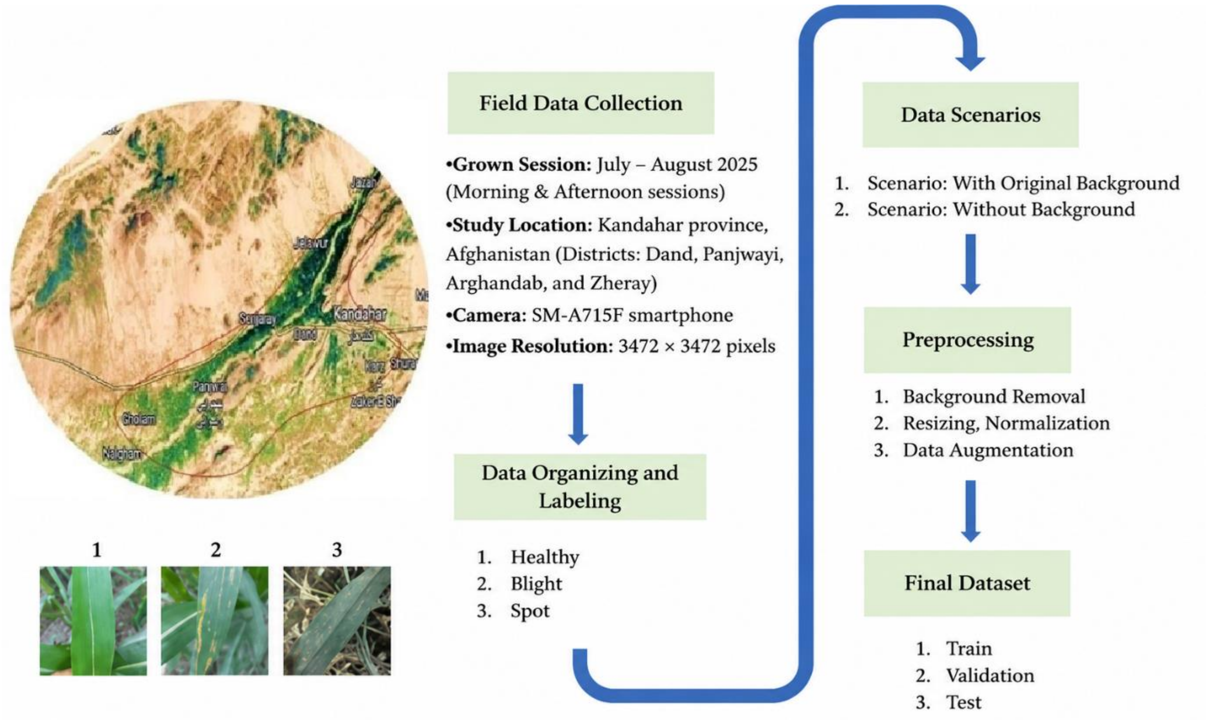

Fig. 1. Workflow of Field Dataset collection and study area map of Kandahar Province showing the surveyed districts

The dataset was organised into training, validation, and testing subsets, as summarised in Table 1, with class labels automatically inferred from the directory structure. In this study, all deep learning models were trained and evaluated exclusively on raw field images containing natural background information. This design choice was intended to reflect real-world agricultural conditions, where maize leaves are observed within complex environmental contexts. Input images were resized according to model-specific requirements to ensure architectural compatibility, with transformer-based and MaxViT models using a resolution of 224 × 224 pixels, while EfficientNet variants employed higher input resolutions consistent with their compound scaling strategy. This unified dataset configuration ensured fair comparison across models while preserving the realistic visual variability present in field conditions. Although images were captured under natural field conditions, basic preprocessing was applied to reduce extreme background noise while preserving a realistic environmental context. The dataset was randomly split into training, validation, and test sets. To reduce potential data leakage, images collected from the same field were assigned to the same subset whenever possible. As images were captured from nearby plants within individual fields, any remaining leakage is acknowledged as a limitation of this study. The dataset and code used in this study are publicly available on GitHub[1].

1- https://github.com/RafiullahRafi/MaizeLeafDisease-Kandahar-HybridData-MaxViT

Table 1- Field dataset distribution for training, validation, and test sets

Dataset Type	Class	Train	Valid	Test
Field data with original background	Spot	597	25	20
	Health	465	20	18
	Blight	828	33	39
	Total	1890	78	77

Data Preprocessing and Augmentation

To enhance model generalisation and mitigate overfitting, a systematic data preprocessing and augmentation pipeline was implemented using the "*torchvision.transforms*" module. The original dataset comprised 785 images across three classes: Spot, Healthy, and Blight. These images were divided into 70% training (630 images), 15% validation (78 images), and 15% test (77 images) sets using stratified sampling by class and district to preserve field-level independence and minimise potential data leakage caused by similar backgrounds and illumination conditions within the same fields. To increase the effective training data and improve model robustness, augmentation techniques such as random horizontal flipping and random rotation within ±10 degrees were applied. As a result, the total number of training images increased from 630 original images to 1890 augmented images across all classes. For the validation and test sets, only resizing and normalisation were applied, with no augmentation, to ensure consistent and unbiased evaluation. Input images were resized according to the architectural requirements of each model to ensure compatibility and optimal feature extraction. Specifically, transformer-based and MaxViT models utilised an input resolution of 224×224 pixels, while EfficientNet variants employed higher resolutions (456×456 for B5, 528×528 for B6, and 600×600 for B7) in accordance with their compound scaling design. During training, data augmentation techniques were applied to increase sample diversity, including random horizontal flipping and random rotation within ±10 degrees. The augmented images were subsequently converted to tensors and normalised using either ImageNet statistics (mean = [0.485, 0.456, 0.406], standard deviation = [0.229, 0.224, 0.225]) or a standardised normalisation scheme ([0.5, 0.5, 0.5]), depending on the specific preprocessing pipeline. For the validation set, only resizing and normalisation were applied, with no augmentation introduced, in order to maintain consistency and ensure unbiased performance evaluation. This differential preprocessing strategy allowed the models to learn robust and invariant features during training while preserving the integrity and reliability of validation results.

Models' Architecture

In this study, several state-of-the-art deep learning architectures were employed to classify maize image data, including EfficientNet variants (B5, B6, and B7), Vision Transformer (ViT), and multiple MaxViT models. These architectures were selected to represent convolutional, transformer-based, and hybrid design paradigms, enabling a comprehensive evaluation of their effectiveness in agricultural image classification tasks.

The EfficientNet family (B5, B6, and B7) is based on a compound scaling strategy that uniformly scales network depth, width, and input resolution to achieve improved performance with optimised computational efficiency. These models utilise Mobile Inverted Bottleneck Convolution (MBConv) blocks combined with depthwise separable convolutions to reduce parameter count while preserving feature representation capability. Additionally, Squeeze-and-Excitation (SE) modules are integrated to enhance channel-wise feature recalibration, allowing the network to focus on more informative features. As the model variant increases from B5 to B7, higher input image resolutions and deeper network structures enable more precise extraction of fine-grained visual patterns present in maize images.

The Vision Transformer (ViT) model adopts a transformer-based architecture that processes images as sequences of fixed-size patches. Each input image is divided into non-overlapping patches, which are linearly embedded and augmented with positional encodings to retain spatial information. The core of ViT consists of stacked transformer encoder layers that employ multi-head self-attention mechanisms to model global contextual relationships among image patches. Feedforward neural networks, combined with residual connections and layer normalisation, further enhance feature learning. This global attention-based design allows ViT to effectively capture long-range dependencies within maize images (Figure 2).

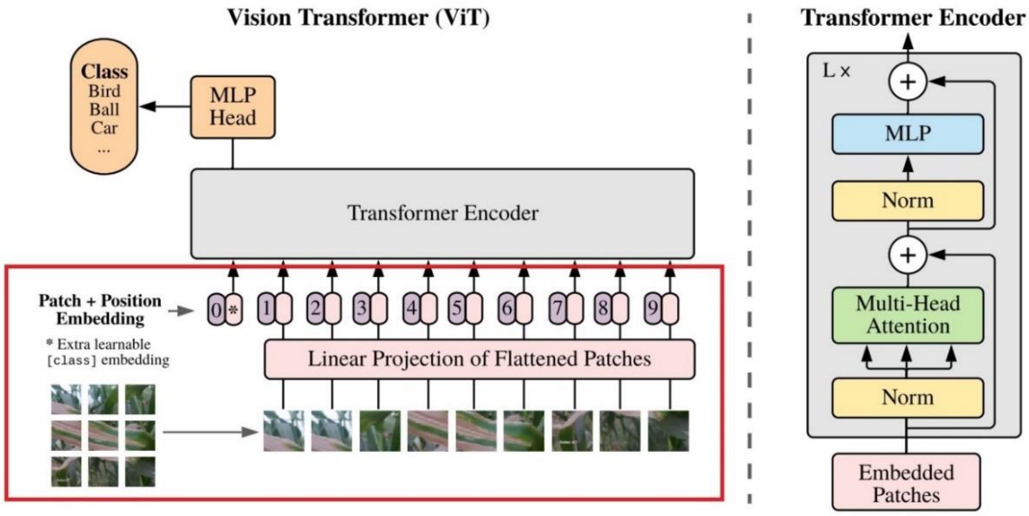

Fig. 2. Architecture of the Vision Transformer (ViT)

The MaxViT architectures represent a hybrid approach that integrates convolutional operations with transformer-based attention mechanisms. Specifically, the MaxViT-Tiny, MaxViT-Small, and MaxViT-Base models combine the strengths of local feature extraction and global context modelling within a unified framework. These models employ MBConv blocks at early stages to capture local spatial information, followed by alternating window-based self-attention and grid-based global attention layers. Window-based attention restricts self-attention computation to localised regions, significantly reducing computational complexity, while grid-based attention enables global feature interaction across the entire feature map. Feedforward networks with normalisation layers are applied to enhance nonlinear feature transformations (Figure 3).

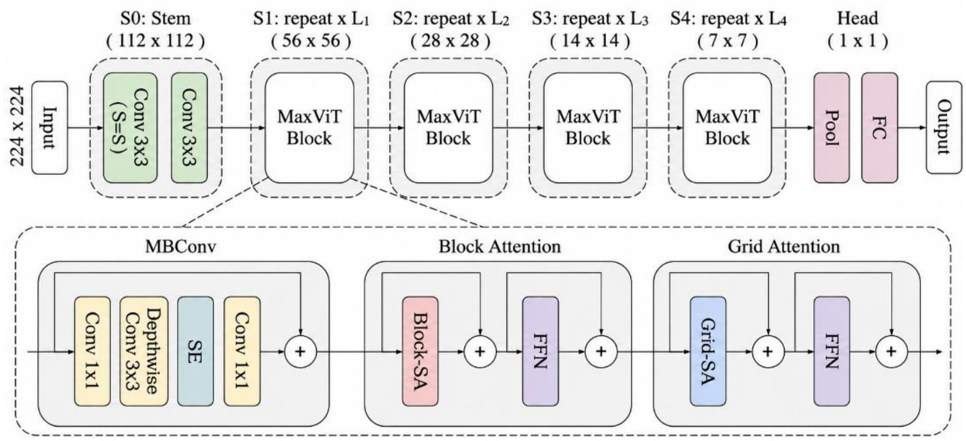

Fig. 3. The MaxViT model architecture combines MBConv blocks with multi-axis attention for efficient feature extraction (Tu *et al.*, 2022)

Overall, the evaluated architectures differ in terms of model capacity, computational cost, and feature learning strategies. Convolutional models such as EfficientNet demonstrate strong performance in capturing local texture and shape features, transformer-based models like ViT excel in modelling global dependencies, while hybrid MaxViT architectures effectively balance both local and global representations. This architectural diversity provides a robust foundation for assessing deep learning performance on maize image classification tasks.

Table 2 summarises the number of parameters, GFLOPs, and deployment suitability of each model used in our experiments. On this basis, MaxViT-Tiny (30.9M parameters, 5.56 GFLOPs) offers the best deployment suitability with estimated mobile inference times of ~150–300 ms, outperforming heavier models like EfficientNet-B7 (66M, 37 GFLOPs) in resource-constrained settings.

Table 2- Overview of model architectures, computational cost (GFLOPs), deployment suitability, and performance metrics

Model	Parameters (M)	GFLOPs	Deployment Suitability
EfficientNet-B5	30	10.0	Moderate – suitable for GPUs or server deployment; relatively heavy for mobile/edge.
EfficientNet-B6	43	19.0	Moderate–Heavy – requires higher GPU memory; less suitable for mobile.
EfficientNet-B7	66	37.0	Heavy – best for high-capacity GPUs/servers; not recommended for edge devices.
Vision Transformer (ViT-Base-224)	86.6	17.0	Heavy – transformer-based; requires GPU/accelerator for efficient deployment.
MaxViT_tiny_rw_224	30.9	5.56	Light–Moderate – suitable for mobile and edge devices with limited resources.
MaxViT_small_tf_224	69	11.7	Moderate – suitable for mid-range GPU/server environments.
MaxViT_base_tf_224	120–135	23–24	Heavy – designed for high-end GPU/server setups; not recommended for edge devices.

Training Configuration

Two training pipelines were followed, depending on the architecture, as shown in Table 3. Transformer-based models (MaxViT_tiny, MaxViT_small, MaxViT_base, and ViT-B/16) were trained using a unified set-up with the *CrossEntropyLoss* function and the AdamW optimiser, while varying weight decay, learning rate, and batch size according to model capacity. All transformer models operated on images resized to 224 × 224 pixels and were trained for 50 epochs, with smaller batch sizes and lower learning rates adopted for larger variants to ensure training stability. In contrast, EfficientNet family-based models (EfficientNet-B5, B6, and B7) followed a second pipeline characterised by higher input image resolutions (ranging from 456 × 456 to 600 × 600), fewer training epochs (20 epochs), and progressively smaller batch sizes as model depth increased. To compensate for memory constraints associated with large input sizes, gradient accumulation was employed for EfficientNet-B6 and B7. The training configurations were tailored to each model family to balance computational efficiency, convergence behaviour, and model scalability.

Overall, hyperparameters were tuned through preliminary validation experiments in accordance with transfer learning best practices for small datasets. Transformers required 50 epochs to achieve stable attention convergence, whereas EfficientNet models converged within 20 epochs due to convolutional efficiency at higher resolutions. To ensure fair and reproducible comparisons, each architectural family was independently optimised while maintaining identical ImageNet pre-training weights, a consistent data augmentation pipeline, and a standardised evaluation protocol across all models.

Also, the training dataset exhibits a moderate class imbalance, with a higher number of blight samples compared to healthy and spot classes. No explicit class re-weighting or resampling strategy was applied in order to reflect realistic field data distributions. This may be a potential bias source toward the majority class.

Table 3- Comparison of training configurations for all utilised models

Parameter	**Transformer-based models**			
	MaxViT_tiny	MaxViT_small	MaxViT_base	ViT-B/16
Loss Function	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss
Optimiser	AdamW	AdamW	AdamW	AdamW
Weight Decay	1e-5	1e-4	1e-4	1e-4
Learning rate	(lr=1e-4)	1e-4	5e-5	3e-5
Image size	224 × 224	224 × 224	224 × 224	224 × 224
Batch Size	32	16	8	16
Epochs	50	50	50	50

Parameter	**EfficientNet family-based models**		
	EfficientNet_b5	EfficientNet_b6	EfficientNet_b7
Loss Function	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss
Optimiser	AdamW	AdamW	AdamW
Weight Decay	1e-5	1e-5	1e-5
Learning rate	1e-4	5e-5	5e-5
Image size	456 × 456	528 × 528	600 × 600
Batch Size	8	4	2
ACCUM_STEPS	1	2	4
Epochs	20	20	20

Evaluation Metrics

To evaluate the performance of the models, a comprehensive set of global and class-specific metrics was utilised, including Categorical Accuracy, Precision, Recall, and F1-score, reported as both macro-averaged and weighted-average values. Furthermore, a detailed per-class analysis was conducted to investigate potential model bias and its generalisation capacity across distinct disease categories. The evaluation framework incorporated Precision, Recall, and an additional metric index, with the respective computational formulas for each metric outlined as follows:

$$\text{Categorical Accuracy} = \text{Number of Correct Predictions} / \text{Total Number of Predictions} \quad (1)$$

$$\text{Macro F1 score} = \frac{1}{c}\sum_{j=1}^{c} F1j \quad (2)$$

$$\text{Weighted F1 score} = \sum_{j=1}^{c} wj \,.\, F1j \quad (3)$$

$$Precicion = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$\text{F1} - \text{Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (6)$$

where TP (True Positives), FP (False Positives), FN (False Negatives); R is the rank position of the model based on the ascending or descending order of the F1-Score, and N is the total number of models; and C is the number of classes.

Results and Discussion

This section presents the experimental results obtained from training and evaluating all deep learning models on a real-world maize leaf dataset collected from Kandahar Province, Afghanistan. All experiments were conducted exclusively using raw field images containing natural background information, in order to reflect realistic agricultural conditions. No background removal or segmentation-based preprocessing was applied in this study. Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and confusion matrix, to provide a comprehensive assessment of classification effectiveness under complex and cluttered field environments. The results demonstrate that the evaluated architectures are capable of learning discriminative features from maize leaf images despite the presence of natural background variations, highlighting their robustness and practical applicability in real-field disease detection scenarios.

Training and Validation Curves

To evaluate the training dynamics, convergence behaviour, and generalisation capability of the models, the training and validation accuracy curves over epochs are presented in Figure 4. The figure combines subplots for all evaluated architectures, allowing direct comparison of their learning patterns on the real-field maize disease dataset.

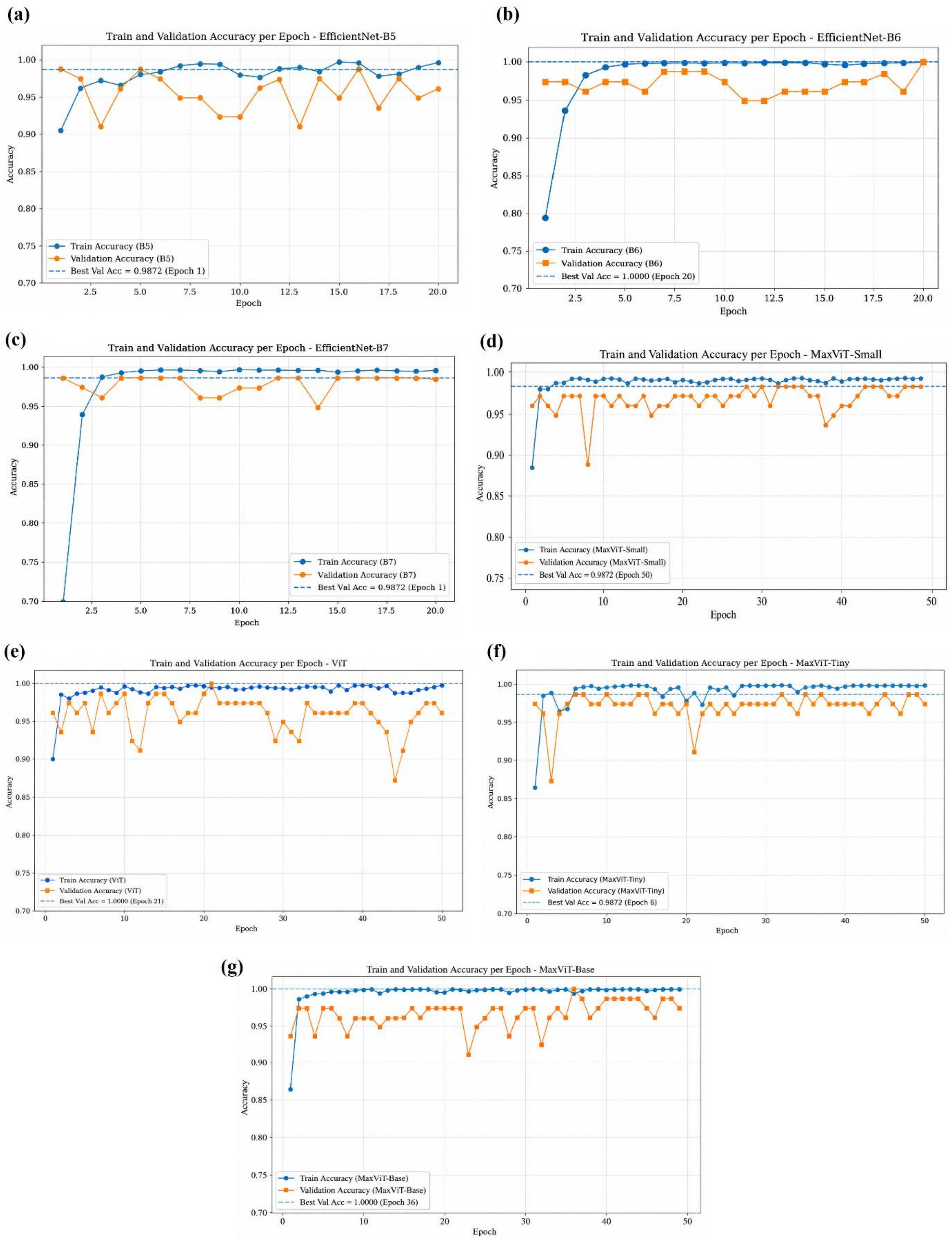

Fig. 4. Training and validation accuracy curves per epoch for the evaluated models: (a) EfficientNet-B5, (b) EfficientNet-B6, (c) EfficientNet-B7, (d) MaxViT-Small, (e) Vision Transformer (ViT), (f) MaxViT-Tiny, and (g) MaxViT-Base. Solid lines represent training accuracy, while dashed lines indicate validation accuracy

The curves reveal stable and consistent convergence across most models, with validation accuracy closely tracking or even occasionally surpassing training accuracy in later epochs, a strong indicator of effective generalisation and minimal overfitting. Specifically:

- EfficientNet family (subfigures a–c): The models exhibit rapid initial convergence, particularly EfficientNet-B7 (c), which reaches near-peak validation accuracy (~0.97) within the first 5–10 epochs and maintains remarkable stability thereafter. Minor fluctuations in validation accuracy are observed but remain within a narrow band (~±0.02), consistent with the benefits of compound scaling and regularisation inherent in EfficientNet architectures.
- Transformer-based and hybrid models (subfigures e–g): Vision Transformer (e) and MaxViT-Tiny (f) show slightly slower but steady improvement, achieving stable validation performance above 0.96 after 20–30 epochs. MaxViT-Tiny demonstrates efficient learning with the smallest gap between training and validation curves among transformer variants, highlighting the advantage of its hybrid multi-axis attention mechanism on limited, real-world data.
- Larger models (subfigures d and g): MaxViT-Small (d) and MaxViT-Base (g) display more pronounced oscillations in validation accuracy, especially in early-to-mid epochs, and exhibit larger final gaps between training and validation curves (particularly MaxViT-Base). This behaviour is indicative of mild overfitting or reduced generalisation when model capacity exceeds the effective information available in the small dataset.

Overall, the absence of significant divergence between training and validation curves across the top-performing models (EfficientNet-B7, ViT, and MaxViT-Tiny) provides robust evidence against substantial overfitting, despite the dataset's limited size, class imbalance, and real-world variability. These patterns corroborate the high test-set performance reported in Table 4 and support the conclusion that the evaluated architectures, when properly configured, are capable of learning discriminative and generalisable features from authentic field-collected maize leaf images.

Performance Evaluation Using Precision, Recall, and F1-Score

Table 4 presents the quantitative performance comparison of the evaluated deep learning models trained on raw maize field images containing natural background information. Model performance was assessed using standard classification metrics, including accuracy (A), precision (P), recall (R), and F1-score, to ensure a comprehensive evaluation of predictive capability and class discrimination. Among the EfficientNet variants, performance improved consistently with increasing model scale and input resolution. EfficientNet-B5 achieved an accuracy of 94%, while EfficientNet-B6 and EfficientNet-B7 demonstrated enhanced performance with accuracies of 96% and 97%, respectively. Notably, EfficientNet-B7 attained an F1-score of 0.9700, reflecting its strong balance between precision (0.9658) and recall (0.9748). This trend highlights the effectiveness of compound scaling in capturing fine-grained visual patterns in maize images under real-field conditions.

The Vision Transformer (ViT) model achieved comparable performance to EfficientNet-B7, with an accuracy of 97% and an identical F1-score of 0.9700. Its high recall value (0.9748) indicates strong sensitivity in correctly identifying maize classes, suggesting that global self-attention mechanisms are effective in modelling contextual dependencies present in complex background environments. Among the hybrid architectures, *MaxViT_tiny_rw_224* demonstrated competitive performance, achieving an accuracy of 97% and an F1-score of 0.9692. This model recorded the highest precision value (0.9741) among all evaluated architectures, indicating robust class discrimination despite its relatively lightweight design. In contrast, MaxViT_small_tf_224 and MaxViT_base_tf_224 exhibited comparatively lower performance, with accuracies of 95% and 92%, respectively. The drop in MaxViT-Base performance is attributed to overfitting on the small dataset, where higher capacity leads to memorisation of noise rather than generalisation. This is supported by larger training-validation accuracy gaps in training curves and increased confusion in the matrix (Table 5).

Table 4- Overall Performance of all models on the original background image data

Model	Batch Size	Image Size	Accumulation Gradient	A	P	R	F1-score
EfficientNet-B5	8	456	1	0.94	0.9251	0.9329	0.9263
EfficientNet-B6	4	528	2	0.96	0.9583	0.9500	0.9512
EfficientNet-B7	2	600	4	0.97	0.9658	0.9748	0.9700
Vision Transformer	16	224	-	0.97	0.9658	0.9748	0.9700
MaxViT_tiny_rw_224	32	224	-	0.97	0.9741	0.9667	0.9692
MaxViT_small_tf_224	16	224	-	0.95	0.9440	0.9333	0.9331
MaxViT_base_tf_224	8	224	-	0.92	0.9023	0.9243	0.9110

A = Accuracy, P = Precision, and R = Recall

Overall, the results indicate that both large-scale convolutional models and transformer-based architectures can achieve high classification accuracy on maize field images with background information. EfficientNet-B7, Vision Transformer, and MaxViT-Tiny emerged as the top-performing models, demonstrating a strong balance between precision and recall. These findings underscore the importance of architectural design choices and input resolution in optimising performance for agricultural image classification tasks under realistic field conditions.

Confusion Matrix Analysis

Table 5 summarises the confusion matrix analysis of all EfficientNet-based and transformer-based models. The confusion matrix analysis provides a detailed insight into the classification behaviour of the evaluated deep learning models for maize leaf disease detection, including Blight, Healthy, and Spot classes. Overall, all models demonstrate strong discriminative capability, particularly in identifying healthy leaves, where perfect classification (100% accuracy) is consistently observed across all architectures without any false positives or false negatives. This indicates that the spectral–textural features of healthy leaves are highly distinguishable from diseased classes in the utilised dataset.

Table 5- Confusion matrices of all models with misclassifications and total images per class

Model	Class	Blight	Healthy	Spot	Misclassifications	Total Images
	Blight	37	1	1	2	39
EfficientNet-B5	Healthy	0	18	0	0	18
	Spot	1	2	17	3	20
	Blight	39	0	0	0	39
EfficientNet-B6	Healthy	0	18	0	0	18
	Spot	1	2	17	3	20
	Blight	38	0	1	1	39
EfficientNet-B7	Healthy	0	18	0	0	18
	Spot	0	1	19	1	20
	Blight	38	0	1	1	39
ViT	Healthy	0	18	0	0	18
	Spot	0	1	19	1	20
	Blight	39	0	0	0	39
MaxViT-Tiny	Healthy	0	18	0	0	18
	Spot	1	1	18	2	20

MaxViT-Small	Blight	39	0	0	0	39
	Healthy	0	18	0	0	18
	Spot	1	3	16	4	20
MaxViT-Base	Blight	36	0	3	3	39
	Healthy	0	18	0	0	18
	Spot	0	3	17	3	20

For the EfficientNet family, EfficientNet-B5 exhibits minor misclassifications, with one Blight sample incorrectly predicted as Healthy and one as Spot, while a small amount of confusion is also observed between Spot and the other classes. EfficientNet-B6 shows improved robustness by correctly classifying all Blight and Healthy samples, with only limited confusion in the Spot class, where a small number of samples are misclassified as Blight and Healthy. EfficientNet-B7 maintains comparable performance, achieving near-perfect recognition, with only one Blight sample confused with Spot and one Spot sample misclassified as Healthy, reflecting a balanced and stable classification behaviour.

The Vision Transformer (ViT) model demonstrates performance similar to EfficientNet-B7, with high accuracy across all classes and minimal confusion occurring only between the Spot and Healthy categories. This suggests that ViT effectively captures global contextual features of maize leaf images, although slight ambiguity remains in visually similar disease patterns.

Among the MaxViT variants, MaxViT-Tiny achieves excellent performance, correctly classifying all Blight and Healthy samples, with only minor misclassification within the Spot class. MaxViT-Small also shows strong results; however, a slightly higher confusion rate is observed for Spot samples being predicted as Healthy, indicating increased sensitivity to inter-class visual similarity. In contrast, MaxViT-Base presents the highest level of confusion among the evaluated models, particularly within the Blight and Spot classes, where several samples are misclassified between these two disease categories. This behaviour may be attributed to increased model complexity, which can lead to overfitting or reduced generalisation when trained on relatively limited datasets.

In summary, the confusion matrix results confirm that all models are highly reliable in identifying healthy maize leaves, while misclassifications predominantly occur between Blight and Spot diseases due to their overlapping visual characteristics. EfficientNet-B6, EfficientNet-B7, ViT, and MaxViT-Tiny demonstrate the most balanced and robust performance, making them suitable candidates for practical deployment in automated maize leaf disease diagnosis systems under real-field conditions.

Qualitative Analysis of Misclassifications

To gain deeper insight into the nature of classification errors and the decision-making behaviour of the models, a qualitative examination of misclassified samples was conducted. Misclassifications were predominantly observed between the Spot and Blight classes, which is consistent with the confusion matrix results presented in Table 5. These two disease categories frequently exhibit overlapping visual characteristics in real-field conditions, including similar lesion colouration (yellow-brown to reddish-brown), irregular lesion shapes, and partial occlusion by shadows, soil particles, or surrounding vegetation.

Selected representative misclassified images from the top-performing models (EfficientNet-B7, Vision Transformer, and MaxViT-Tiny) are shown in Table 5 along with their true and predicted labels. The examples illustrate common sources of confusion: In several cases, early-stage Spot lesions were misclassified as Blight due to subtle boundary diffusion and background interference (e.g. soil texture or lighting gradients), which partially masked fine-grained lesion patterns. Conversely, Blight samples with elongated or coalesced lesions were occasionally predicted as Spot when the model overly focused on localised discolouration rather than the broader lesion distribution.

A small number of Spot samples were incorrectly assigned to the Healthy class, typically when lesions were small, peripheral, or heavily shadowed, reducing their visual saliency. These patterns highlight the challenge of distinguishing visually similar foliar diseases under uncontrolled field variability, where environmental factors (variable illumination, complex backgrounds, and partial occlusions) can degrade discriminative feature extraction.

Table 5. Representative misclassified samples from the test set with true and predicted labels:

- True: Spot, predicted: Blight (EfficientNet-B7) – early-stage small lesions confused by background soil noise.
- True: Blight, predicted: Spot (Vision Transformer) – elongated lesions misinterpreted due to partial shadow occlusion.
- True: Spot, predicted: Healthy (MaxViT-Tiny) – peripheral small lesions suppressed by strong lighting gradients and surrounding healthy tissue.
- True: Blight, predicted: Spot (MaxViT-Tiny) – coalesced lesions mistaken for clustered Spot symptoms.

The qualitative observations align with the quantitative confusion matrix and underscore the strengths of attention-based architectures (ViT and MaxViT-Tiny) in capturing global context, while also revealing persistent limitations in handling subtle inter-class differences under realistic agricultural imaging conditions. Future improvements could incorporate targeted data augmentation strategies (e.g. shadow simulation, lesion boundary enhancement) or multi-scale feature fusion to further reduce such confusions.

Discussion

The results of this study demonstrate that modern deep learning architectures are highly effective for plant disease classification when trained and evaluated on real-field images. Unlike many previous studies that relied on controlled laboratory datasets or images with uniform backgrounds, the dataset used in this research was collected directly from agricultural fields in Kandahar Province, Afghanistan. As a result, the images inherently contain complex and realistic backgrounds, including soil texture, variable illumination, shadows, and surrounding vegetation. This significantly increases the difficulty of the classification task and provides a more reliable assessment of model robustness under real-world conditions. Importantly, this real-field variability allows interpreting architectural behaviour under environmental complexity, rather than merely reporting accuracy values obtained from simplified datasets.

Among the evaluated models, EfficientNet variants exhibited a clear performance trend, with accuracy and F1-score improving as the network depth and input image resolution increased. EfficientNet-B7 achieved one of the highest performances, reaching an accuracy of 97% and an F1-score of 0.9700. This superior performance can be attributed to compound scaling, which proportionally increases depth, width, and resolution, enabling finer representation of lesion morphology and subtle texture differences between Spot and Blight in cluttered scenes. These findings align with prior research highlighting the effectiveness of compound scaling in EfficientNet architectures for agricultural image analysis (Amin *et al.*, 2022; Rajeena *et al.*, 2023). However, unlike previous maize disease studies conducted on curated datasets (e.g. PlantVillage), our results demonstrate that such CNN scaling remains effective under uncontrolled field variability. However, it is important to note that higher-performing EfficientNet variants require smaller batch sizes and gradient accumulation due to increased computational demands, which may limit their deployment in resource-constrained settings.

Transformer-based models also demonstrated competitive and, in some cases, superior performance. The Vision Transformer (ViT) achieved an accuracy of 97% and an F1-score comparable to EfficientNet-B7, despite using significantly smaller input images (224×224). This suggests that global self-attention mechanisms effectively capture long-range contextual relationships, allowing the model to focus on diagnostically relevant leaf regions while suppressing irrelevant background noise, an advantage particularly relevant in real-field images with occlusion and heterogeneous backgrounds. Similarly, the MaxViT_tiny_rw_224 model achieved the highest precision (0.9741) and a strong F1-score (0.9692), highlighting the advantage of hybrid architectures that integrate convolutional inductive biases with transformer-based global attention. The combination of local texture extraction (convolution) and multi-axis global attention contributed to reduced confusion between visually similar classes, particularly Spot and Blight, demonstrating the benefit of hybrid inductive biases in complex scenes.

In contrast, larger MaxViT configurations, such as MaxViT_base_tf_224, showed reduced performance, with an accuracy of 92% and an F1-score of 0.9110. This performance drop indicates that increased architectural capacity does not necessarily improve generalisation on moderately sized, heterogeneous datasets and may introduce overfitting risk. These observations are consistent with previous findings that lightweight or moderately scaled transformer models often generalise better in practical agricultural applications (Nishankar *et al.*, 2025; Ullah *et al.*, 2024).

When compared with classical machine learning approaches and earlier CNN-based systems, the proposed results indicate a substantial performance improvement. Traditional methods relying on handcrafted features and classifiers such as SVMs struggled to handle environmental variability and achieved lower generalisation performance in field conditions (Ren, Zhang, & Wang, 2019). In contrast, the deep learning models evaluated in this study effectively learned hierarchical and discriminative representations directly from raw images, enabling robust disease classification despite background noise and real-world distortions. This confirms earlier conclusions that deep learning architectures have become the dominant paradigm in agricultural image analysis (Attallah, 2023; Naderi Beni, Bagherpour, & Parian, 2024). However, the present study extends prior maize disease research by systematically comparing CNN, transformer, and hybrid architectures under authentic field conditions, thereby evaluating architectural robustness rather than confirming high accuracy on controlled datasets.

Overall, the findings of this study validate the applicability of both CNN-based and transformer-based models for real-world plant disease detection in challenging agricultural environments. The use of authentic field images from Kandahar Province enhances the ecological validity of the results and supports the deployment of such models in practical decision-support systems for farmers. Future work may focus on optimising lightweight hybrid architectures to balance accuracy and computational efficiency, as well as extending the dataset to cover additional crops, disease stages, and seasonal variations.

Limitation

This study, while providing ecologically valid insights from real-field conditions in Kandahar Province, has several limitations that should be acknowledged. First, the dataset size is relatively small, particularly the test set (total n = 77 images: 20 Spot, 18 Healthy, and 39 Blight). Such limited test samples reduce the statistical power of the reported metrics and widen confidence intervals, potentially leading to over-optimistic performance estimates. Small test sets are a common challenge in region-specific agricultural studies due to labelling costs and seasonal constraints, but future work should aim to expand the test set to at least 100–200 images per class. Second, a moderate class imbalance exists in the training set (Blight: 828, Healthy: 465, and Spot: 597 before augmentation), which may bias models toward the majority class (Blight). Although no explicit re-weighting or oversampling was applied to preserve the natural field distribution, this remains a potential source of bias, especially for minority classes like Healthy and Spot. Weighted loss functions or focal loss could be explored in follow-up studies.

Third, data were collected exclusively from four districts in Kandahar Province during July–August 2025 using a single smartphone model. This geographic and temporal restriction limits generalisation to other maize varieties, growth stages, soil types, climatic zones, or imaging devices prevalent in Afghanistan or elsewhere. Smartphone-based capture introduces realistic variability but may not fully represent professional camera quality or drone imagery used in large-scale precision agriculture. Finally, although field-level stratified splitting minimised data leakage from identical plants or nearby rows, some residual correlation between images within the same field may persist, slightly inflating performance. Cross-field or leave-one-field-out validation would provide a more rigorous assessment in future iterations.

Conclusion

This study systematically evaluated state-of-the-art deep learning architectures namely EfficientNet variants, Vision Transformer (ViT), and MaxViT models, for maize leaf disease classification using authentic real-field images from Kandahar Province, Afghanistan. By preserving natural backgrounds, variable illumination, and complex environmental conditions, the work addresses a critical gap in prior studies that rely predominantly on controlled datasets. Results demonstrate robust performance under realistic conditions, with EfficientNet-B7, ViT, and MaxViT-Tiny achieving 97% accuracy and F1-scores ≥ 0.97. MaxViT-Tiny stands out for its superior precision (0.9741) and favourable trade-off, offering high discriminative power with only 30.9 million parameters and 5.56 GFLOPs, enabling estimated mobile inference times of ~150–300 ms, highly suitable for resource-constrained precision agriculture settings. These findings confirm the effectiveness of hybrid attention mechanisms in handling field variability and complex backgrounds, while highlighting that increased model capacity (e.g. MaxViT-Base) does not always improve generalisation on limited real-field data. Minor misclassifications, primarily between visually similar Spot and Blight classes, underscore persistent challenges in distinguishing subtle symptom differences. Despite promising results, limitations include the small test set size (n = 77), geographic restriction to Kandahar Province, and potential residual intra-field correlation, which may slightly inflate performance and limit broader generalisability. Future work should priorities: (1) expanding the dataset across multiple seasons, regions, and maize varieties; (2) integrating explainable AI techniques (e.g. Grad-CAM) for deeper misclassification analysis; and (3) optimising lightweight models for edge deployment to support real-time, on-farm decision-making in resource-limited environments.

Author Contributions

R. Rafi: Data acquisition, Data pre- and post- processing, Statistical analysis, Validation, Visualisation, Text mining

A. Soleimanipour: Supervision, Conceptualisation, Methodology, Software services, Review and editing services

A. Rezaei Asl: Supervision, Technical advice, Review and editing services

References

1. Ahmadi, I. (2025). Detection and Classification of Some Diseases of Tomato Crops Using Transfer Learning. *Journal of Agricultural Machinery*, *15*(3), 319-335. https://doi.org/10.22067/jam.2024.88500.1258
2. Alshammari, H., Gasmi, K., Ltaifa, I. Ben, Krichen, M., Ammar, L. Ben, & Mahmood, M. A.

(2022). Olive Disease Classification Based on Vision Transformer and CNN Models. *Computational Intelligence and Neuroscience*, *2022*, 1-11. https://doi.org/10.1155/2022/3998193
3. Amin, H., Darwish, A., Hassanien, A. E., & Soliman, M. (2022). End-to-End Deep Learning Model for Corn Leaf Disease Classification. *IEEE Access*, *10*, 31103-31115. https://doi.org/10.1109/ACCESS.2022.3159678
4. Attallah, O. (2023). Tomato leaf disease classification via compact convolutional neural networks with transfer learning and feature selection. *Horticulturae*, *9*(2), 149. https://doi.org/10.3390/horticulturae9020149
5. Bruns, H. A. (2017). Southern Corn Leaf Blight: A Story Worth Retelling. *Agronomy Journal*, *109*(4), 1218-1224. https://doi.org/10.2134/agronj2017.01.0006
6. Chang, J., Wei, S., Liu, Y., Zhao, Z., & Shi, J. (2025). Harnessing Genetic Resistance in Maize and Integrated Rust Management Strategies to Combat Southern Corn Rust. *Journal of Fungi*, *11*(1), 41. https://doi.org/10.3390/jof11010041
7. Dai, D., Xia, P., Zhu, Z., & Che, H. (2023). MTDL-EPDCLD: A Multi-Task Deep-Learning-Based System for Enhanced Precision Detection and Diagnosis of Corn Leaf Diseases. *Plants*, *12*(13), 2433. https://doi.org/10.3390/plants12132433
8. Dhaka, V. S., Meena, S. V., Rani, G., Sinwar, D., Kavita, K., Ijaz, M. F., & Woźniak, M. (2021). A Survey of Deep Convolutional Neural Networks Applied for Prediction of Plant Leaf Diseases. *Sensors*, *21*(14), 4749. https://doi.org/10.3390/s21144749
9. El Sakka, M., Mothe, J., & Ivanovici, M. (2024). Images and CNN applications in smart agriculture. In *European Journal of Remote Sensing*, *57*(1), 1-30. Taylor & Francis. https://doi.org/10.1080/22797254.2024.2352386
10. Eli-Chukwu, N. C. (2019). Applications of Artificial Intelligence in Agriculture: A Review. *Engineering, Technology & Applied Science Research*, *9*(4), 4377-4383. https://doi.org/10.48084/etasr.2756
11. Entuni, C. J., & Zulcaffle, T. M. A. (2022). Identification of Corn Leaf Diseases Comprising of Blight, Grey Spot and Rust Using DenseNet-201. *Borneo Journal of Resource Science and Technology*, *12*(1), 125-134. https://doi.org/10.33736/bjrst.4224.2022
12. Fraiwan, M., Faouri, E., & Khasawneh, N. (2022). Classification of Corn Diseases from Leaf Images Using Deep Transfer Learning. *Plants*, *11*(20), 1-14. https://doi.org/10.3390/plants11202668
13. Idress, K. A. D., Gadalla, O. A. A., Öztekin, Y. B., & Baitu, G. P. (2024). Machine Learning-based for Automatic Detection of Corn-Plant Diseases Using Image Processing. *Tarım Bilimleri Dergisi*, *30*(3), 464-476. https://doi.org/10.15832/ankutbd.1288298
14. Jain, S., Sahni, R., Khargonkar, T., Gupta, H., Verma, O. P., Sharma, T. K., Bhardwaj, T., Agarwal, S., & Kim, H. (2022). Automatic Rice Disease Detection and Assistance Framework Using Deep Learning and a Chatbot. *Electronics (Switzerland)*, *11*(14), 2110. https://doi.org/10.3390/electronics11142110
15. Jaiswal, M., Nagwanshi, K. K., Kumar, R., Gaurav, S., & Watti, Y. (2024). Classification of Plant Disease using a State-of-the Art Deep learning Algorithm on a Tesla GPU. *International Journal of Computing*, 240-246. https://doi.org/10.47839/ijc.23.2.3542
16. Keerthivasan, M., Kokilavani, S., Shanthi, M., Dheebakaran, G., Selvi, R. P., Murugan, M., ..., & Kumar, M. S. (2026). Machine Learning for Detection of Pests in Tomato: A Review. *Journal of Agricultural Machinery*, *16*(1), 223-244. https://doi.org/10.22067/jam.2025.90762.1314
17. Koroshi Talab, H., Mohammad Zamani, D., & Gholami Par-Shokohi, M. (2025). Diagnosis and Classification of Two Common Potato Leaf Diseases (Early Blight and Late Blight) Using Image Processing and Machine Learning. *Journal of Agricultural Machinery*, *15*(1), 65-79.

https://doi.org/10.22067/jam.2024.86571.1226

18. Ma, Z., Wang, Y., Zhang, T., Wang, H., Jia, Y., Gao, R., & Su, Z. (2022). Maize leaf disease identification using deep transfer convolutional neural networks. *International Journal of Agricultural and Biological Engineering*, *15*(5), 187-195. https://doi.org/10.25165/j.ijabe.20221505.6658
19. Masood, M., Nawaz, M., Nazir, T., Javed, A., Alkanhel, R., Elmannai, H., Dhahbi, S., & Bourouis, S. (2023). MaizeNet: A Deep Learning Approach for Effective Recognition of Maize Plant Leaf Diseases. *IEEE Access*, *11*, 52862-52876. https://doi.org/10.1109/ACCESS.2023.3280260
20. Naderi Beni, A. N., Bagherpour, H., & Parian, J. A. (2024). Development and Optimization of a Novel Deep Learning Model for Diagnosis of Quince Leaf Diseases. *Journal of Agricultural Machinery*, *14*(4), 445-458. https://doi.org/10.22067/jam.2024.88013.1248
21. Nishankar, S., Pavindran, V., Mithuran, T., Nimishan, S., Thuseethan, S., & Sebastian, Y. (2025). ViT-RoT: Vision Transformer-Based Robust Framework for Tomato Leaf Disease Recognition. *AgriEngineering*, *7*(6), 185. https://doi.org/10.3390/agriengineering7060185
22. Pacal, I., & Işık, G. (2025). Utilizing convolutional neural networks and vision transformers for precise corn leaf disease identification. *Neural Computing and Applications*, *37*(4), 2479-2496. https://doi.org/10.1007/s00521-024-10769-z
23. Panigrahi, K. P., Das, H., Sahoo, A. K., & Moharana, S. C. (2020). Maize Leaf Disease Detection and Classification Using Machine Learning Algorithms. *Advances in Intelligent Systems and Computing*, *1119*(January), 659-669. https://doi.org/10.1007/978-981-15-2414-1_66
24. Pranta, A. S. U. K., Fardin, H., Debnath, J., Hossain, A., Sakib, A. H., Ahmed, M. R., Haque, R., Reza, A. W., & Dewan, M. A. A. (2025). A Novel MaxViT Model for Accelerated and Precise Soybean Leaf and Seed Disease Identification. *Computers*, *14*(5), 197. https://doi.org/10.3390/computers14050197
25. Rajeena, P. P. F., Su, A., Moustafa, M. A., & Ali, M. A. S. (2023). Detecting Plant Disease in Corn Leaf Using EfficientNet Architecture—An Analytical Approach. *Electronics*, *12*(8), 1938. https://doi.org/10.3390/electronics12081938
26. Rashid, R., Aslam, W., Aziz, R., & Aldehim, G. (2024). An Early and Smart Detection of Corn Plant Leaf Diseases Using IoT and Deep Learning Multi-Models. *IEEE Access*, *12*(January), 23149-23162. https://doi.org/10.1109/ACCESS.2024.3357099
27. Ren, T., Zhang, Y., & Wang, C. (2019, September). Identification of corn leaf disease based on image processing. In *2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)* (pp. 165-168). IEEE. https://doi.org/10.1109/ICISCAE48440.2019.221610
28. Setiawan, W., Rochman, E. M. S., Satoto, B. D., & Rachmad, A. (2022). Machine Learning and Deep Learning for Maize Leaf Disease Classification: A Review. *Journal of Physics: Conference Series*, *2406*(1), 1-11. https://doi.org/10.1088/1742-6596/2406/1/012019
29. Shamloo, R., Soleimanipour, A., & Rezaei Asl, A. (2025). Biosystems Engineering and Renewable Energies Weeds detection in saffron fields using an improved YOLO v5 model. *Biosystems Engineering and Renewable Energies*, *1*(1), 44-50.
30. Sopialena, S., Suyadi, S., & Noor, S. A. (2022). Ecosystem Monitoring on Leaves of Leaf Rust Disease of Maize (*Zea mays* L.). *Caraka Tani: Journal of Sustainable Agriculture*, *37*(1), 89. https://doi.org/10.20961/carakatani.v37i1.34920
31. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., & Li, Y. (2022). MaxViT: Multi-axis Vision Transformer. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 13684 LNCS* (pp. 459–479). Springer Science and Business Media Deutschland GmbH.

https://doi.org/10.1007/978-3-031-20053-3_27

32. Ullah, W., Javed, K., Khan, M. A., Alghayadh, F. Y., Bhatt, M. W., Al Naimi, I. S., & Ofori, I. (2024). Efficient identification and classification of apple leaf diseases using lightweight vision transformer (ViT). *Discover Sustainability*, *5*(1), 116. https://doi.org/10.1007/s43621-024-00307-1
33. Waheed, A., Goyal, M., Gupta, D., Khanna, A., Hassanien, A. E., & Pandey, H. M. (2020). An optimized dense convolutional neural network model for disease recognition and classification in corn leaf. *Computers and Electronics in Agriculture*, *175*, 105456. https://doi.org/10.1016/j.compag.2020.105456
34. Yang, S., Yao, J., & Teng, G. (2024). Corn Leaf Spot Disease Recognition Based on Improved YOLOv8. *Agriculture (Switzerland)*, *14*(5), 666. https://doi.org/10.3390/agriculture14050666

ارزیابی مدل‌های یادگیری عمیق مبتنی بر مبدل برای طبقه‌بندی بیماری‌های برگی ذرت در شرایط واقعی مزرعه: مطالعه موردی استان قندهار، افغانستان

رفیع‌اله رفیع[۱]، علیرضا سلیمانی‌پور[۲*]، عباس رضایی اصل[۲]

تاریخ دریافت: ۱۴۰۴/۱۰/۲۹
تاریخ پذیرش: ۱۴۰۵/۰۱/۲۹

چکیده

شناسایی دقیق و به‌موقع بیماری برگ ذرت (*Zea mays* L.) برای بهبود بهره‌وری محصول و اطمینان از امنیت غذایی امری ضروری است. این مطالعه به بررسی عملکرد مدل‌های یادگیری عمیق مبتنی بر مبدل در شناسایی خودکار بیماری‌های برگ ذرت با استفاده از یک مجموعه داده جمع‌آوری‌شده از مزارع ذرت استان قندهار افغانستان پرداخته است. مجموعه داده تحت شرایط طبیعی مزرعه جمع‌آوری شد و پس زمینه پیچیده برگ، نورپردازی متغیر در طول روز و جهت‌گیری مختلف برگ باعث شد تا محیط واقعی مزرعه حفظ شود. چندین معماری پیشرفته یادگیری عمیق مبتنی بر مبدل مورد ارزیابی قرار گرفت. این معماری‌ها شامل مبدل بینایی (ViT)، مبدل بینایی چندمحوره مرکب (Max-Vit-Tiny، MaxViT-small و MaxViT-base) بودند. نتایج با برخی مدل‌های سنگین یاگیری عمیق مبتنی بر شبکه کانولوشنی EfficientNet (شامل نسخه‌های B5، B6 و B7) مقایسه شدند. نتایج نشان داد که مدل‌های مبتنی بر مبدل، همانند مدل‌های مبتنی بر CNN، عملکردی قابل‌اتکا (دقت بالای ۹۷٪) در محیط‌های چالش‌برانگیز مزرعه‌ای دارند. در بین مدل‌های ارزیابی‌شده، مدل EfficientNet-B7، ViT و MaxViT_tiny با صحت تشخیص ۹۷٪ بهترین عملکرد را داشتند، درحالی‌که مدل MaxVit_tiny بالاترین دقت را ثبت کرد (۰/۹۷۴۱). این مورد نشان‌دهنده ظرفیت عمومی‌سازی بالای این مدل‌هاست. خطاهای تشخیص عمدتا بین آن‌دسته از کلاس‌های بیماری مشاهده شد که شباهت بصری زیادی با هم داشتند. درحالی‌که دقت و حساسیت کلی مدل‌ها به‌طورپیوسته بالا بود. این یافته‌ها اثرگذاری معماری‌های مرکب و مبتنی بر توجه در کابردهای شرایط واقعی برای شناسایی بیماری ذرت را تایید می‌کند و نشان‌دهنده پایایی آن‌ها در سامانه‌های کشاورزی دقیق است.

واژه‌های کلیدی: EfficientNet، Max-ViT، Vit، بیماری برگ ذرت، داده مزرعه

۱- گروه مهندسی کشاورزی، دانشکده علوم گیاهی، پوهنتون ملی علوم کشاورزی و تکنولوژی افغانستان (ANASTU)، قندهار، افغانستان
۲- گروه مهندسی مکانیک بیوسیستم، دانشکده آب و خاک، دانشگاه علوم کشاورزی و منابع طبیعی گرگان، گرگان، ایران
(* - نویسنده مسئول: Email: asoleimani@gau.ac.ir)

https://doi.org/10.22067/jam.2026.97474.1465