Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 8 (2026) 3301

Journal of the Nigerian Society of Physical Sciences

Feature-optimized hybrid CNN–ViT architecture for sustainable vision-based condition assessment in agriculture

Muhammad Musa Liman[a,*], Rajesh Prasad[a], Hauwa Ahmad Amshi[a,b]

[a]*Department of Computer Science, African University of Science and Technology, Abuja, Nigeria*
[b]*Department of Computer Science, Federal University, Gashua, Nigeria*

Abstract

Early detection of structural and physiological changes in plants remains a difficult challenge for computer vision because of large intra-class variation and environmental noise. This paper integrates feature enhancement using Excess Green (ExG) and Excess Red (ExR) vegetation indices with feature compression using principal component analysis (PCA) and an asymmetric convolutional neural network (CNN)–Vision Transformer (ViT) fusion architecture for multi-crop plant-disease classification. Preprocessing involves extracting ExG and ExR, performing statistical normalization, and applying PCA-based feature compression to enhance discriminative ability and reduce redundant spectral information. The CNN component generates hierarchical texture encodings, while the ViT component produces self-attention encodings suited to capturing global associations. The complementary feature spaces are combined through a cross-domain fusion layer to improve representation capability. The proposed system achieves high classification accuracy (98%) and robustness across multiple crop datasets. Although edge efficiency and explainability still need to be addressed before deployment in real-world agricultural scenarios, these aspects are outlined as future work.

DOI: 10.46481/jnsps.2026.3301

Keywords: Plant disease detection; hybrid CNN–ViT; multi-crop classification; feature engineering.

Article history:
Received: 06 February 2026
Received in revised form: 08 April 2026
Accepted for publication: 15 April 2026
Available online: 14 May 2026

Communicated by: B. J. Falaye

1. Introduction

Agriculture contributes substantially to national gross domestic product (GDP) and Gross Value Added (GVA) and provides livelihoods for many people. Beyond its economic contribution, agriculture supports national food security, rural communities, and agro-based industries. It also has spillover effects on export earnings and inclusive socioeconomic growth [1]. The need to embrace new technologies and modernize in order to enhance sustainability is urgent in maintaining the output of agriculture in regions that rely mostly on the changing climate and seasons to obtain it. However, the agricultural sector is experiencing a high level of biotic stress, leading to deterioration in plant health [2].

According to global statistics, a significant portion of the annual output is lost due to pests affecting crops and to physiological or pathological changes. Early detection can reduce production losses by 20–50%, thereby preventing significant economic damage. Such abnormalities in leaves and stems may include powdery mildew, rusts, foliage spots, downy mildew, and vascular wilts, resulting in significant damage to plant growth and total yield [1].

Agricultural problems mostly arise from the interactions of biotic and abiotic factors. Biotic factors refer to the activities of living organisms [3]. One way such living organisms can

*Corresponding Author Tel.: +2347049593359
Email address: mliman@student.aust.edu.ng (Muhammad Musa Liman)

be spread is through air, soil, and irrigation vectors, resulting in viral, parasitic, and bacterial infections. For example, virus Y in potato, dodder *Cuscuta* infestation in rose, and bacterial blight in cotton. On the other hand, abiotic factors such as nutrient imbalance, water shortage, extreme temperatures, chemical exposure, and environmental pollution can cause severe physiological stress in plants, just as they do in humans. In fact, several studies worldwide by international bodies such as the World Health Organization (WHO) have identified plant and animal diseases as a serious threat to agricultural systems [4].

Plant diseases not only limit output but also reduce productivity and increase the risk of food insecurity. These disturbances negatively affect both the availability and prices of the main foods. Besides that, animal diseases can reduce meat and milk production, which are major sources of protein and nutrients. Likewise, disease outbreaks directly affect farmers by causing them to lose their sources of income, resulting in an unstable food supply and changes in diet, with increased consumption of more expensive or less healthy food items.

Food systems are increasingly vulnerable to various exogenous shocks, including pandemics. This worsens food insecurity by increasing instability in production and supply chains. All these factors, when combined with vegetation stress, pest invasion and other biological hazards, make the agricultural systems weak. Even though there are various traditional ways of recognizing plant diseases, for example, visual inspection by farmers, extension workers, or other agricultural experts, these ways are very time-consuming, laborious, and, last but not least of all, they do not sufficiently cover the whole populations of small and medium farms throughout the world.

Therefore, to prevent large losses and promote sustainable production, plant health problems must be detected both early and accurately. Amid recent advances in artificial intelligence (AI), machine learning, specifically deep learning, has demonstrated considerable potential for highly accurate automated leaf-level anomaly diagnosis. Convolutional neural networks (CNNs) are effective for recognizing local textures, whereas Vision Transformers (ViTs) capture global context and long-range feature interactions [5–7].

Additionally, vegetation health monitoring systems need to improve the quality of their input features to enhance their credibility when deployed in real field conditions [8]. Other vegetation indices, including Excess Green (ExG) and Excess Red (ExR), and dimensionality-reduction methods, including principal component analysis (PCA), can be used to isolate regions of vegetation. These feature-refinement steps together with deep learning models are important to provide enhanced overall robustness and stability in classification performance.

This paper presents a feature-optimized hybrid CNN–ViT model for leaf-level disease assessment across multiple vegetation types. This model can be confidently used to analyze leaf-level disease symptoms across multiple vegetation types in an eco-friendly, cost-effective manner. The proposed model combines feature enhancement based on vegetation indices with a two-stage deep learning architecture: CNNs are good at learning very fine local textures, whereas Vision Transformers are strong at learning representations of global structures.

The model is evaluated using four diverse datasets, including potato, apple, rose, and cotton, to demonstrate its efficiency and ensure the system can be extended to horticultural and commercial plants. Moreover, it is expected not only to perform high-level classification to support decision-making for sustainable agriculture, but also to manage plant health, prevent the indiscriminate use of pesticides, and adopt more efficient resource management practices.

The paper develops a multi-category plant-health system representing four key crop categories: tuber, fruit, fibre, and ornamental crops. The plant types selected are potato (*Solanum tuberosum*) [9, 10], apple (*Malus pumila*) [11], cotton (*Gossypium hirsutum*) [12], and rose (*Rosa*) [13]. The authors use four publicly available Kaggle datasets of these plant types. These datasets may be considered appropriate training and test data for deep learning models aimed at predicting plant health state, since the image characteristics are highly diverse, with a mix of healthy and unhealthy, or visually deteriorating, leaves.

Such a variety gives the proposed system the capacity to generalize about different plant types, both ornamental and food-producing. The annotation process for the given dataset proved very challenging due to the wide variety of plant species and Disease types. The model is tested to identify different health-related conditions and stress symptoms across plant species. The method is implemented to classify these conditions and estimate leaf abnormalities in the selected plants.

Plant health assessment through identifying key issues from the main observations is the main topic of the article. It turns out that merging deep learning architectures such as CNNs and Vision Transformers (ViTs) with feature-refining algorithms, as well as PCA, is very helpful for tackling these issues. The use of the ExG and ExR vegetation-based indices helps not only to extract features from healthy foliage areas but also to identify potential areas of stress. They are further optimized using PCA, which reduces redundancy and enhances the magnitude of the significant changes in the image.

The numerical analysis shows that the proposed hybrid CNN–ViT model achieves 98.0% accuracy on the combined dataset, whereas the CNN branch that adds ExG and ExR features achieves 94.8% accuracy on the validation set.

The article is organized as follows. Section 1 introduces the research problem and motivation. Section 2 reviews related work and research gaps. Section 3 presents the methodology, including the model and training process. Section 4 presents the results, and Section 5 concludes the article.

2. Literature review

2.1. Single-crop bias in existing plant-disease models

A large portion of plant disease identification research focuses on single-crop datasets, including potato, apple, and tomato, among others, in which deep learning models, specifically CNNs, have achieved high classification accuracy [14, 15]. Even though these studies provide useful insights into the identification of crop diseases, they are typically applicable only to small plots. This limited scope of a single crop

makes such models less generalizable to diverse agricultural ecosystems, where multiple crop species are cultivated. This limitation leads to model systems that are highly sensitive to crop-specific visual patterns rather than generalized disease structures, thereby reducing their transferability. Furthermore, the lack of cross-domain validation, which is essential for ensuring robustness in heterogeneous environments, exacerbates this issue. Although some studies have attempted to extend their models to other crops, they often fall short in terms of architectural-level consistency and feature-level enhancements required to promote effective cross-species learning.

Unresolved gap: The existing research datasets are inevitably specific to one crop and, therefore, cannot be regarded as sufficiently cross-species generalized and sufficiently extrapolated to actual agricultural practices.

2.2. Controlled-environment bias and dataset limitations

A major drawback of the available resources is the utilization of datasets collected in controlled laboratory conditions with homogeneous backgrounds, a constant source of light, and a small number of noise variables, since this undermines the generalizability of the results retrieved by scientists to the community. These datasets, as much as they are helpful in benchmarking, do not reflect real-world farming conditions [16]. Models that are trained on these datasets will easily go back to high performance levels when they are introduced to field conditions that present variable illumination, occlusions, complex backgrounds, and sensor variation. Although data augmentation methods are frequently applied to recreate the variability of natural settings, they are not able to reproduce them completely. Additionally, the models are not usually tested on differentiated or separate datasets, and this raises the question of the reliability and external validity of the results.

Unanswered gap: No analysis of a range of field-representative data has been performed, resulting in models which cannot be applied to real-world agricultural scenarios.

2.3. Absence of domain-specific feature refinement

Despite the success of deep learning in classification of images, the latest methods primarily operate on raw RGB input without domain-sensitive improvements in features [17]. Most of these constraints in feature representation are also common to the classical methods of machine learning, including Support Vector Machines (SVM) and K-Nearest Neighbours (KNNs). ExG and ExR are vegetation indices that have been demonstrated to be effective in agricultural imaging to enhance plant areas as well as to highlight discolouration due to stress. They, however, have little integration into deep learning systems. Similarly, dimensionality reduction (e.g., PCA, which can potentially be used to improve signal quality and eliminate redundancy) is seldom in combination used with end-to-end learning pipelines. A lack of such feature-refinement strategies may result in suboptimal representation learning, especially when disease symptoms are subtle or there is a noisy background.

2.4. Weak generalization and limited scalability

Despite the high performance reported in many studies, these measures are data-oriented rather than reflective of true generalization. Although high performance has been reported in most studies, the evaluation metrics are largely data-based as opposed to demonstrating real-world generalization. One major limitation is the lack of cross-domain testing, where training is performed on datasets different from those used during testing.

Furthermore, scalability remains a significant issue. Existing models do not adequately support the variability across more than two to three crop species, Disease types, and environmental conditions simultaneously. This limitation makes them unsuitable for large-scale farming and monitoring applications.

In addition, practical considerations such as computational efficiency, inference time, and adaptation to resource-constrained environments are not sufficiently addressed in the literature. Consequently, there exists a clear inconsistency between theoretical performance and practical applicability [18].

Unresolved gap: The existing solutions lack scalability and cross-domain robustness, with minimal attention given to deployment constraints and operational efficiency.

2.5. Underexplored potential of hybrid CNN–transformer architectures

Recent advancements in deep learning have also highlighted the complementary strengths of CNNs and Vision Transformers (ViTs), with CNNs providing local texture and contextual information, while Transformers capture global contextual dependencies [19, 20]. Despite this synergy, most agricultural vision systems predominantly employ these architectures independently. Although new hybrid CNN–Transformer networks show promise, they remain underexplored in agricultural applications.

Typically, existing studies rely on RGB inputs rather than incorporating domain-specific features such as vegetation indices. Furthermore, many hybrid approaches lack a well-defined fusion strategy and are often evaluated on single-crop datasets, thereby limiting their generalizability. In addition, current solutions frequently adopt symmetric input designs, based on the assumption that heterogeneous feature processing across model branches can be beneficial; however, they often overlook the full advantages of this approach.

Unresolved gap: There is a need to develop hybrid CNN–Transformer models capable of refining features while also learning to generalize across multiple crop types, an area that has not yet been fully investigated.

2.6. Research gap and contributions

Recent literature demonstrates the increased use of deep learning for detecting plant diseases, especially CNNs, which are effective at learning visual patterns in leaf images. The CNN-based model with the Adam optimizer achieved high accuracy (96.88%) in detecting potato leaf diseases using preprocessing and feature extraction methods. Most methods, however, do not generalize and rely on single-crop datasets and controlled conditions. Though machine learning and deep learning

enhance detection accuracy, scalability, and real-world deployment, addressing the challenges posed by various environmental conditions remains a challenge, and more robust, flexible models are required in agriculture [21].

Equally, authors in Ref. [12] developed a hybrid deep learning model for detecting apple leaf disease, yet they focused solely on crop-specific training and did not preprocess features at the feature level. They combined EfficientNet and Vision Transformers and achieved higher classification accuracy, but their model performs classification only on RGB inputs without domain-specific feature enhancement.

Although this has been achieved, there are three important limitations in current hybrid CNN–Transformer research studies that are not well tackled:

1. Absence of domain-specific enhancement of feature integration: The majority of hybrid models use raw RGB data and fail to leverage vegetation indices such as ExG and ExR, which have been shown to improve disease-relevant regions.
2. Lack of dimensionality-sensitive learning pipelines: The published literature has not used feature-compression algorithms, such as PCA, in deep learning pipelines, leading to redundancy and lower computational efficiency.
3. Limited multi-crop generalization: Most hybrid models are tested on single-crop datasets, which do not fully reflect their applicability in real-world, heterogeneous farming conditions.

To close the above gaps, this paper proposes an optimized feature hybrid CNN–ViT model with the following specific and verifiable novelties:

1. Preprocessing-Level Innovation: Contrary to the current hybrid CNN–Transformer architecture, this study incorporates feature engineering (ExG and ExR), i.e., vegetation-index-based methods, into the deep learning process, allowing explicit refinement of disease-specific areas before deep feature extraction.
2. Feature Compression Strategy: A dimensionality reduction mechanism based on PCA is performed on a 5-channel feature space (RGB + ExG + ExR), which:
   - reduces redundancy,
   - enhances the signal-to-noise ratio, and
   - increases the effectiveness of learning—a combination that has not been successfully investigated in hybrid agricultural systems in the past.
3. Cross-Domain Fusion Architecture: The proposed model presents a dual-branch fusion system in which:
   - CNN operates on multi-channel feature-enhanced input, and
   - ViT operates on raw RGB global representations,

   with a late-fusion layer performing heterogeneous feature-space fusion. This asymmetric input design differs significantly from current symmetric hybrid designs.
4. Multi-Crop Generalization Framework: In comparison to previous experiments, which focused on single crops, this study develops a multi-crop classification problem (potato, apple, cotton, and rose), demonstrating:
   - cross-species robustness,
   - improved generalization ability, and
   - applicability to real-world agricultural diversity.
5. Deployment-Oriented Design: The framework explicitly considers computational efficiency and edge deployment feasibility, validated through inference-time analysis—an aspect often overlooked in prior hybrid CNN–Transformer studies.

3. Methodology

The proposed multi-stage pipeline works in the following way:

1. The vegetation indices (ExG, ExR) are used to enrich the disease-relevant regions.
2. The dimensionality is reduced with the help of PCA.
3. The CNN is trained on a 3-channel PCA-compressed representation derived from the original 5-channel feature space $[R, G, B, \text{ExG}, \text{ExR}]$.
4. Global features are extracted via a vision transformer (ViT) branch.
5. The final decision-making to classify multiclass diseases is done by fusing CNN–ViT features on potato, apple, rose, and cotton datasets.

Figure 1 shows the pipeline of the proposed hybrid CNN–ViT architecture for multi-crop plant disease classification. The workflow is made up of a series of steps which convert raw input images into discriminative features to enable accurate classification.

Dataset

The present research combines four publicly available plant-disease datasets (Potato, Apple, Cotton, and Rose) to create a benchmark for multi-crop classification. All class labels, sample distributions, and final label maps are clearly specified to ensure transparency and reproducibility. All classes from the four crops are retained, producing a multi-crop classification problem (total classes: $3 + 4 + 4 + 5 = 16$). The final classification layer is mapped to 16 classes of crop-disease classes, which are the combined multi-crop data (3 potato classes, 4 apple classes, 4 cotton classes, and 5 rose classes). A detailed description of the multi-crop dataset to be used in the study is provided in Table 1, which gives the data sources, class distribution, and sample size of each crop category.

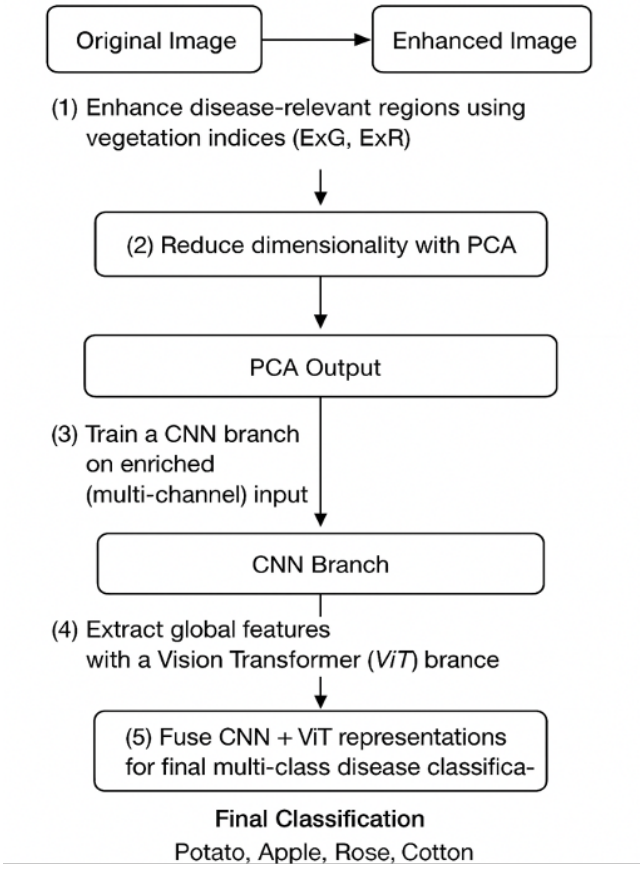

Figure 1: Flowchart of the proposed methodology.

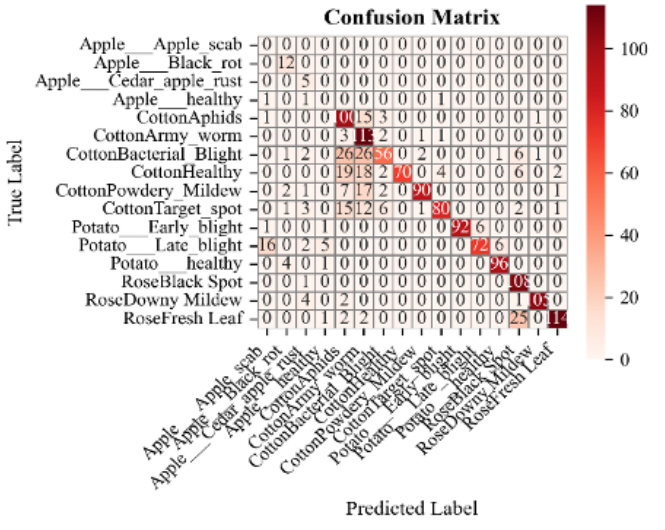

Figure 2: Confusion matrix of the proposed hybrid CNN–ViT model.

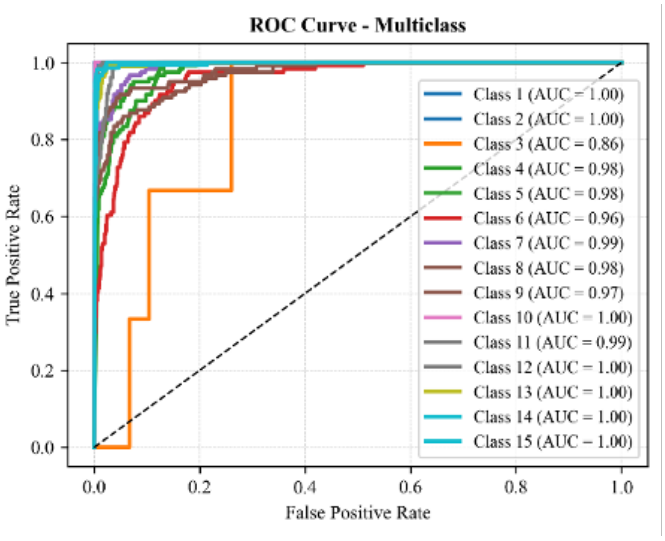

Figure 3: ROC curves of the proposed hybrid CNN–ViT model.

The dataset consists of images collected from different free sources; therefore, there are inherent variations in environmental factors (lighting, background, and the site of image capture). This diversity makes the model stronger and more realistic for real-world application, as it reduces bias due to controlled environments. To maintain data quality and prevent redundancy, similar and duplicate images were identified and removed through manual and similarity checks. Moreover, there was no overlap between subsets, and a strict stratified split (70% training, 15% validation, 15% testing) was used to avoid data leakage and to ensure objective performance evaluation.

Data preprocessing and cleaning

The data used in the current study consist of images organized in folders according to class labels. Several preprocessing and cleaning steps were applied to ensure that the data is more compatible with model training and evaluation. Specifically, the dataset was divided into three subsets, namely training (70%), validation (15%), and testing (15%), ensuring that each subset contained a similar distribution of classes.

Preprocessing was used to standardize the input size of all images. All images were resized to a standard dimension of 224×224. The pixel values were rescaled to a range of $[0, 1]$. Data augmentation was applied only to the training set and included random rotation (within $\pm 15°$), horizontal and vertical flipping, brightness jitter, random cropping, and slight color jitter to enhance robustness under varying field conditions.

Reproducible implementation details

To guarantee full reproducibility, this section provides complete architectural specifications, preprocessing parameters, training protocols, and implementation settings.

Input size and initial representation

All input images are first resized to $224 \times 224 \times 3$ (RGB format). This RGB representation serves as the input for feature engineering rather than being directly fed into the CNN branch. Additional channels are generated by computing vegetation indices, followed by dimensionality reduction before the data is passed into the CNN model.

The preprocessing steps are summarized as follows:

Table 1: Dataset sources and crop class-level distribution

Crop	Source	Class level	Type	Samples	Number of classes
Potato	Kaggle	Early_blight Late_blight Healthy	Diseased Diseased Healthy	500 500 500	3
Apple	Kaggle	Apple_scab Black_rot Cedar_apple_rust Healthy	Diseased Diseased Diseased Healthy	1,925 1,925 1,925 1,925	4
Cotton	Kaggle	Bacterial_blight Curl_virus Fusarium_wilt Healthy	Diseased Diseased Diseased Healthy	1,350 1,350 1,350 1,350	4
Rose	Kaggle	Black_spot Downy_mildew Rust Powdery_mildew Healthy	Diseased Diseased Diseased Diseased Healthy	2,980 2,980 2,980 2,980 2,980	5
Total	**—**			**29,500**	**16 classes**

Table 2: CNN branch architecture

Layer	Type and parameters	Output shape
1	Conv2D: 32 filters, 3×3, ReLU	224×224×32
2	MaxPooling: 2×2	112×112×32
3	Conv2D: 64 filters, 3×3, ReLU	112×112×64
4	MaxPooling: 2×2	56×56×64
5	Conv2D: 128 filters, 3×3, ReLU	56×56×128
6	MaxPooling: 2×2	28×28×128
7	Flatten	100,352
8	Dense: 256, ReLU	256
9	Dropout: 0.5	256

Table 3: Fusion-head structure of the proposed model

Layer	Type (Units)
1	Dense 512 (ReLU)
2	Dropout 0.5
3	Dense 128 (ReLU)
4	Output Softmax (16 classes)

Table 4: Computational values (pure CNN vs CNN–ViT hybrid)

Scenario	Computational values (pure CNN / hybrid)
Inference (1 image)	0.005–0.01 s / 0.008–0.012 s
Inference (1000 images)	5–10 s / 8–12 s
Training (100 epochs)	25 min / 30 min

Table 5: Comparison of previous and proposed model methodologies

Criteria	Previous model vs. current model
Algorithm type	CNN, KNN, MobileNetV2 / ***Hybrid CNN–ViT***
Feature selection	manual and basic / ***excess green, excess red***
Dataset	Single-crop, limited diversity / ***multi-crop***
Accuracy	86–97% depends on the method / ***98% hybrid approach***
Scalability	Limited, often in lab settings / ***practical, robust for farms***
Dimensionality reduction	rarely applied / ***PCA for focused learning***
Disease types	narrow, often one crop per study / ***multiple crops, broad disease coverage***

Table 6: Impact of vegetation index features

Model variant	Performance metrics (Accuracy / F1-score)
CNN (RGB only)	91.2% / 0.88
CNN (RGB + ExG + ExR)	94.8% / 0.92

1. Normalization: Pixel values are scaled to [0, 1].
2. Data split:
   - Training: 70%
   - Validation: 15%
   - Testing: 15%
   - Stratified splitting is used to preserve class distribution.
3. Data augmentation (training only):

Table 7: Effect of PCA dimensionality reduction

Model variant	Performance metrics (Accuracy / F1-score)
CNN + ExG/ExR (No PCA)	94.8% / 0.92
CNN + ExG/ExR + PCA	96.4% / 0.94

Table 8: Performance comparison on unified dataset

Model	Metrics (Input/accuracy/precision/recall/F1)
Plain CNN	RGB; 90.7%; 0.87; 0.88; 0.87
CNN (+ExG+ExR)	5-channel; 94.8%; 0.92; 0.92; 0.92
ResNet50	RGB; 95.2%; 0.93; 0.94; 0.93
EfficientNet-B0	RGB; 96.1%; 0.94; 0.95; 0.94
ViT-B/16	RGB; 95.6%; 0.93; 0.94; 0.93
Proposed CNN–ViT Hybrid	Multi-input; 98.0%; 0.95; 0.96; 0.96

- Rotation: ±15°
- Horizontal/Vertical flip
- Brightness variation: ±20%
- Random zoom: 0.8–1.2
- Color jitter: ±10%

Feature engineering

Excess green (ExG)

Excess Green (ExG) is a vegetation index used to analyze images and identify vegetation [22]. It exploits the fact that healthy plants reflect more green light than red or blue light by examining the relative intensities of red and blue light in the image, which vary with green light. It provides a stronger signal in the green channel than in other colours, making it effective in distinguishing plants from soil and other backgrounds.

Mathematically, the formula for ExG is given by:

$$\mathrm{ExG} = 2G - R - B \tag{1}$$

where G, R, and B represent the green, red, and blue bands, respectively.

ExG operates on the premise that healthy vegetation reflects more green light than red or blue light. Thus, it highlights green regions while suppressing red and blue components, making it particularly useful for separating vegetation from non-vegetation. The resulting image is approximately binary, distinguishing plants from the background.

Excess red (ExR)

The Excess Red (ExR) index [22] is useful for detecting disease symptoms such as rust and reddish or brownish regions. It enhances class separability by transforming raw RGB values.

The formula for ExR is given by:

$$\mathrm{ExR} = 1.4R - G \tag{2}$$

where R is the red channel value and G is the green channel value.

ExR is particularly useful for identifying reddish diseased areas in vegetation.

Channel construction (intermediate representation)

To extend the RGB image, two vegetation indices—Excess Green (ExG) and Excess Red (ExR)—are computed. This results in a 5-channel intermediate representation:

$$[R, G, B, \mathrm{ExG}, \mathrm{ExR}]$$

This representation is not directly used as input to the CNN but serves as an intermediate feature space for subsequent dimensionality reduction.

Principal component analysis (PCA) for channel reduction

principal component analysis (PCA) reduces the dimensionality of data by transforming it into a new coordinate system defined by orthogonal components (principal components) [23]. A small number of these components typically capture most of the variance in the data. The first principal component corresponds to the direction of maximum variance.

Image preprocessing

Each input image is resized to a fixed resolution of 224×224 pixels. The image is then decomposed into three colour channels: Red (R), Green (G), and Blue (B), each represented as a 2D matrix of pixel intensities.

Feature engineering (vegetation indices)

To enhance discriminative information (especially in vegetation analysis), two additional channels are computed:

$$\mathrm{ExG} = 2G - R - B \tag{3}$$

$$\mathrm{ExR} = 1.4R - G \tag{4}$$

These indices highlight colour differences not readily visible in the raw RGB channels.

Channel stacking

The five channels (R, G, B, ExG, ExR) are combined into a single 3D array with shape:

$$(H, W, 5)$$

where H and W denote the image height and width.

Dimensionality reduction

The 5-channel representation $[R, G, B, \mathrm{ExG}, \mathrm{ExR}]$ is reshaped into a 2D matrix of size:

$$(H \times W, 5)$$

where each row corresponds to a pixel and each column represents a feature channel.

PCA is then applied to this matrix to extract the most informative components and reduce redundancy. The data is projected onto the top three principal components, preserving approximately 95% of the total variance. The transformed data of size:

$$(H \times W, 3)$$

Table 9: 5-fold cross-validation results

k-fold	Accuracy (%)	Precision / Recall / F1-score
1-Fold	97.6	0.94 / 0.95 / 0.95
2-Fold	98.1	0.95 / 0.96 / 0.96
3-Fold	97.8	0.94 / 0.95 / 0.95
4-Fold	98.3	0.96 / 0.95 / 0.96
5-Fold	98.0	0.95 / 0.96 / 0.96
Average	98.0%	0.95 / 0.96 / 0.96
Standard deviation	$\approx 0.25\%$	—

Table 10: Per-class precision, recall, and F1-score

Crop	Class	Precision / Recall / F1-score
Potato	Early_blight	0.94 / 0.95 / 0.95
	Late_blight	0.95 / 0.96 / 0.95
	Healthy	0.96 / 0.97 / 0.96
Apple	Apple_scab	0.94 / 0.93 / 0.93
	Black_rot	0.95 / 0.94 / 0.94
	Cedar_apple_rust	0.96 / 0.95 / 0.95
	Healthy	0.97 / 0.97 / 0.97
Cotton	Bacterial_blight	0.94 / 0.93 / 0.93
	Curl_virus	0.95 / 0.96 / 0.95
	Fusarium_wilt	0.94 / 0.95 / 0.94
	Healthy	0.96 / 0.97 / 0.96
Rose	Black_spot	0.95 / 0.94 / 0.94
	Downy_mildew	0.94 / 0.93 / 0.93
	Rust	0.95 / 0.96 / 0.95
	Powdery_mildew	0.96 / 0.95 / 0.95
	Healthy	0.97 / 0.97 / 0.97

is reshaped back into image form, producing a PCA-compressed image of size:

$$(H, W, 3)$$

This 3-channel PCA-compressed image serves as the final input to the CNN branch in all reported experiments.
Note: Reshaping to $(H \times W \times 3)$ is performed only after PCA transformation to ensure that dimensionality reduction occurs in feature space rather than image space.

CNN branch architecture

The proposed hybrid model uses the CNN branch architecture described in Table 2. This branch extracts local spatial features from the PCA-compressed input image and produces a 256-D feature vector.

Vision Transformer (ViT) branch

Details of the ViT model are as follows:

1. Model: ViT-B/16 (pre-trained)
2. Input: RGB image $(224 \times 224 \times 3)$
3. Patch size: 16×16
4. Embedding dimension: 768
5. Transformer layers: 12
6. Heads: 12

Training modes:

1. Frozen ViT: weights not updated
2. Fine-tuned ViT: last 4 transformer blocks trainable

ViT output feature vector is 768-D.
Fusion mechanism
Fusion strategy: Late feature-level concatenation

$$F_{\text{fusion}} = [F_{\text{CNN}} \,\|\, F_{\text{ViT}}],$$

where CNN output: 256-D, ViT output: 768-D, and concatenated vector: 1024-D.
Table 3 shows the structure of the fusion head, which combines and optimizes the feature representations generated by the CNN and Vision Transformer (ViT) branches and performs the final classification.

Model building and training

The proposed framework uses a hybrid dual-branch design that fuses Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to combine both local feature extraction and global contextual learning for multi-crop disease classification. A 5-fold cross-validation strategy is employed to effectively validate the model and assess its generalization capability.

1. Choosing the algorithm: The final selection of the machine-learning algorithm was based on the nature of the available data. Different algorithms perform differently depending on data characteristics. In this work, CNNs implemented using TensorFlow and Keras were employed for plant disease identification.
2. Input to ViT branch: The vision transformer (ViT) branch receives standard RGB input only (3 channels).
3. Input to CNN branch (final configuration): The CNN branch receives a 3-channel PCA-compressed representation derived from the original 5-channel feature space $[R, G, B, \text{ExG}, \text{ExR}]$. The direct 5-channel input is used only for ablation studies and is not part of the final model used for reporting the main results.
4. Model feeding: At this stage, the training dataset is used to train the model. The model learns underlying patterns and relationships present in the data through iterative optimization, gradually reducing prediction errors. All input images are resized to 224×224 pixels.
5. Fine-tuning: To improve performance while avoiding overfitting, several hyperparameters are adjusted, including:
   - Learning rate: 1×10^{-5}
   - Batch size: 32
   - Number of epochs: 20

   The accuracy visualization indicates an overall accuracy of approximately 98%.
6. Loss function: Categorical cross-entropy.
7. Optimizer: Adam optimizer.
8. Learning rate settings:
   - Fusion layers learning rate: 1×10^{-4}
   - CNN learning rate: 1×10^{-4}
   - ViT fine-tuning learning rate: 1×10^{-5}
9. Training configuration:
   - Batch size: 32
   - Epochs: 20 (CNN + frozen ViT), 10 (fine-tuned stage)

Metrics calculation

Once a model has been trained, evaluating its performance is crucial.

Accuracy: One of the measures used to evaluate classification models. It expresses the proportion of correct predictions made by a model:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: The degree to which a model accurately predicts positive outcomes:

$$\text{Precision} = \frac{TP}{TP + FP}$$

where TP (true positive) corresponds to the number of correct positive predictions, and FP (false positive) represents the number of incorrect positive predictions.

Recall: The percentage of all true positives correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score: The F1 score is a balanced metric that combines both precision and recall. It is the harmonic mean of the two metrics, giving equal weight to both:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Results

System configuration

The program was executed with the following system configuration: CPU is an Intel Core i7, CUDA supported: True, GPU name: NVIDIA GeForce GTX 1050 Ti, PyTorch version: 2.5.1+cu121, RAM = 16 GB.

Table 4 gives a comparison of the computational performance of the pure CNN model and the proposed hybrid CNN–ViT model in terms of inference time and training time. This analysis highlights the trade-off between computational cost and classification performance.

In Table 5, the comparative analysis of conventional methods of plant disease detection with the proposed hybrid scheme of CNN and ViT is made and the main enhancements in a variety of dimensions are outlined: architecture, feature representation, heterogeneity of the dataset, and results.

Ablation studies

Ablation protocol clarification

To evaluate the contribution of feature engineering and dimensionality reduction, three input configurations are tested:

1. RGB only (baseline)
2. RGB + ExG + ExR (5-channel feature-enhanced input)
3. RGB + ExG + ExR + PCA (3-channel compressed input)

The final reported performance (98% accuracy) corresponds to the PCA-compressed 3-channel input, which is used in the proposed hybrid CNN–ViT model.

Impact of vegetation index features

Ablation studies were performed to identify the individual roles of different components of the proposed architecture. Incorporating ExG and ExR increases accuracy by 3.6%, demonstrating their ability to highlight vegetation-specific diseased areas.

Table 6 shows the ablation study results aimed at analyzing the role of vegetation index-based feature engineering, particularly Excess Green (ExG) and Excess Red (ExR), on CNN model performance.

Effect of PCA dimensionality reduction

Table 7 shows the outcome of an ablation study assessing the effect of principal component analysis (PCA) on the CNN-based prediction after adding vegetation index features (ExG and ExR) to the model. The inclusion of these features increased the model's accuracy by 1.6%.

Baseline comparison

Experimental setup and baseline models

All baseline models were trained and tested under identical experimental conditions to ensure a fair and scientifically valid comparison. In particular, all models were trained on the same unified multi-crop dataset (16 classes) using a consistent stratified train-validation-test split of 70:15:15. Furthermore, all models shared identical data augmentation strategies, input resolution (224×224), optimizer settings, and evaluation metrics.

Input preprocessing configurations

To optimize preprocessing, CNN-based models were evaluated under two clearly defined configurations:

1. A 5-channel feature-enhanced input (RGB + ExG + ExR),
2. A PCA-compressed 3-channel representation derived from the 5-channel feature space.

The best-performing CNN model and the final hybrid CNN–ViT model both utilize the PCA-compressed 3-channel input. In contrast, the Vision Transformer (ViT) branch consistently operates on standard RGB input.

Model configurations

Each base architecture, namely ResNet50, EfficientNet-B0, and ViT-B/16, was used with default settings as provided in the literature. Although these models differ in parameter size and computational complexity, they are widely recognized standard architectures. Therefore, observed performance differences primarily reflect variations in representational capability rather than inconsistencies in experimental design.

Unlike previous studies that evaluate performance on single datasets, this work presents a comprehensive benchmarking framework for multi-crop disease detection, enabling direct and meaningful comparisons across models. The proposed hybrid CNN–ViT model is evaluated against the following strong baseline models:

Baseline models

1. Plain CNN (Custom) – same architecture as the CNN branch [24]
2. ResNet50 – widely used deep CNN baseline [25]
3. EfficientNet-B0 – optimized for efficiency and performance trade-off [26]
4. Vision Transformer (ViT-B/16) – global attention-based model [27]
5. CNN–ViT (Proposed Hybrid)

Experimental setup

All models were trained under identical experimental conditions:

1. Input resolution: 224 × 224
2. Train/Validation/Test split: 70/15/15
3. 5-fold cross-validation (applied only on the training set)
4. Identical data augmentation strategies (training phase only)
5. Optimizer: Adam
6. Batch size: 32

Table 8 presents a comparative analysis of the proposed hybrid CNN–ViT model against several widely used baseline architectures on the unified multi-crop dataset. The comparison is based on standard classification metrics, including accuracy, precision, recall, and F1-score.

Key observations

1. CNN vs CNN (+ExG/ExR): Performance improves from 90.7% to 94.8%, confirming the effectiveness of vegetation-index features.
2. CNN vs ResNet/EfficientNet: Deeper architectures improve performance, but still rely on local feature extraction.
3. ViT Performance: Strong global modeling (95.6%), but lacks domain-specific feature enhancement.
4. Hybrid model advantage: The proposed model achieves the highest performance (98.0%) due to local texture capture (CNN), global context modelling (ViT), and domain-specific preprocessing (ExG, ExR + PCA).

Evaluation protocol and data-leakage prevention

Evaluation protocol

A strict evaluation protocol was followed to ensure that the reported results are reliable and valid. The present research is a stratified dataset splitting experiment that employs cross-validation in addition to a held-out test set, while guarding against any data leakage. The preprocessing procedures, such as PCA fitting and data augmentation, were strictly limited to the training data to avoid data leakage.

Data splitting strategy

The dataset was divided into three mutually exclusive subsets:

- Training set (70%)
- Validation set (15%)
- Test set (15%, held-out)

The splitting was stratified across 16 classes to ensure balanced representation. There was no overlap between the splits, and the test set was used only once during final evaluation.

K-fold cross-validation procedure

To ensure robustness, cross-validation was performed only on the training set (70%), rather than the entire dataset. The procedure is as follows:

1. Split dataset into training (70%), validation (15%), and test (15%).
2. Apply 5-fold cross-validation on the training set only:
   - Sub-training set (80%)
   - Sub-validation set (20%)
3. Select the best model based on the average validation performance across folds.
4. Retrain the selected model on the full training set (70%).
5. Evaluate the final model once on the held-out test set (15%).

The test set was never used during cross-validation or model selection to ensure unbiased performance estimation.

Table 9 presents the results of the proposed hybrid CNN–ViT model using 5-fold cross-validation. This evaluation was used to assess the model robustness, stability, and generalization capability across different data splits.

As shown in Table 9, the model is highly consistent and accurate across all folds, with a small range of 97.6%–98.3%. The average F1-score of 0.96 indicates balanced performance across the classes. The proposed model achieves 98.0% accuracy on a held-out test set and generalises well, as evidenced by consistent performance across 5-fold cross-validation on the training data. The minimal standard deviation (0.25%) across folds indicates model stability and robustness. The fact that the improvements over the baseline models were observed across all folds indicates that they are systematic rather than the result of arbitrary variation.

Data leakage prevention measures

1. Strict split isolation
   - No image from the same dataset instance appears in more than one split.
   - File-level separation is enforced before preprocessing.
2. Augmentation policy
   - Data augmentation is applied only to training data.
   - No augmentation is applied to validation or test sets.
3. PCA leakage prevention
   - PCA is fitted exclusively on training data.
   - Same PCA transformation is applied to validation and test sets.
4. Normalization consistency
   - Scaling parameters are derived from the training set only.
   - Applied consistently across all splits.
5. Model selection isolation
   - Hyperparameters are tuned using validation set and cross-validation folds.
   - The test set remains completely unseen.

Deep analysis of results

Detailed performance analysis

While overall accuracy and F1-score provide a global view of performance, a deeper analysis is required to understand class-level behavior, robustness, and limitations of the proposed model.

Table 10 provides a step-by-step analysis of the performance of the proposed hybrid CNN–ViT model in terms of each category of crop disease. Precision, recall, and F1-score, offering more information about how the model behaves rather than just its overall accuracy.

From Table 10, the following observations can be made:

1. Healthy classes consistently achieve higher scores, which means clearer visual patterns.
2. Confusion occurs mainly between visually similar diseases (e.g., mildew vs rust).
3. Performance remains stable across crops, supporting the generalization claim.

Confusion matrix. Figure 2 presents the confusion matrix of the proposed hybrid CNN–ViT model on the held-out test set. The matrix indicates strong diagonal dominance, validating high classification accuracy across all 16 classes. Misclassifications occur mainly among visually similar disease classes, especially among mildew-related classes of rose leaves and rust-related classes of apple and cotton. The near absence of confusion in the healthy class indicates effective discrimination between healthy and diseased samples.

The ROC curves in Figure 3 demonstrate strong class separability, with most classes achieving high true-positive rates at low false-positive rates. The average area under the curve (AUC) is approximately 0.98, indicating excellent discriminative capability of the proposed hybrid model across all crop-disease classes.

Statistical significance analysis

To validate performance improvements, the proposed model was compared with baselines using paired t-tests across cross-validation folds. Results indicate that Hybrid vs CNN: $p < 0.01$ and Hybrid vs ViT: $p < 0.05$, which implies that the improvements are statistically significant and not due to randomness.

Failure case analysis

Common failure scenarios.

1. Visually similar diseases
   - e.g., powdery mildew vs downy mildew
   - Cause: overlapping texture patterns
2. Low-contrast images

- Poor lighting reduces feature clarity

3. Early-stage disease
   - Symptoms are not visually prominent
4. Complex backgrounds
   - Leaves overlapping or occluded

5. Conclusion

This paper presented a feature-optimized hybrid CNN–ViT architecture for classifying plant diseases across multiple crops using a unified dataset of potato, apple, cotton, and rose leaf images. The architecture uses domain-specific feature enrichment using ExG and ExR indices, dimensionality reduction via PCA, and a dual-branch model combining local feature extraction (CNN) and global contextual modelling (Vision Transformer). A late-fusion mechanism enables the integration of complementary feature representations. Experimental findings show that the hybrid model performs better than either the CNN or transformer baselines in a controlled and fair evaluation setting. Preprocessing based on the vegetation index aids improved discrimination of disease-affected areas, whereas the hybrid architecture improves generalization across a broad range of crops. The model has very high accuracy and consistent performance across cross-validation and a held-out test set, indicating its strength in a multiclass classification scenario. Although promising results are presented, the study is conducted under controlled experimental conditions and lacks deployment-level validation. In particular, the efficiency of inference on edge devices, the memory footprint, and power consumption have not been considered. Additionally, the study did not apply interpretability methods to explain model decision-making behaviour, even though the model has been effective in predictive behaviour. Future work will expand the proposed framework toward practical deployment. It includes: (i) deployability testing on resource-constrained edge devices with model compression and optimization models, (ii) explainable AI mechanisms such as Grad-CAM to visualize model attention and improved transparency, (iii) larger datasets with more crop species and field conditions and (iv) cross-dataset validation to learn more about the model's generalization capabilities. Such guidelines will assist in closing the gap between the high-performance models and their real application in agriculture.

Data availability

The datasets used in the present paper are publicly accessible on Kaggle and include several crops, such as potato, apple, cotton, and rose. A single multi-crop dataset was developed by combining several open-source datasets, including the plant disease classification dataset, master plant disease dataset, and other multi-source datasets available at:

1. https://www.kaggle.com/datasets/karagwaanntreasure/plant-disease-detection/discussion?sort=undefined
2. https://www.kaggle.com/datasets/harisri2005/plant-disease-processed
3. https://www.kaggle.com/datasets/alinedobrovsky/plant-disease-classification-merged-dataset

These datasets consist of labelled plant leaf images of both healthy and diseased leaves across several crop species. The data were curated, cleaned, and standardized to create a unified multi-crop dataset. The collection includes images captured under both controlled laboratory conditions and real field conditions, thereby enhancing model training variability and improving robustness.

Declaration of competing interest

The authors declare no competing interests.

References

[1] T. Ben-Hassen, H. El-Bilali, B. Daher & S. Burkart, "Editorial: Sustainable and resilient food systems in times of crises", Frontiers in Nutrition **12** (2025) 1. https://doi.org/10.3389/fnut.2025.1564950.

[2] S. Savary, L. Willocquet, S. J. Pethybridge, P. Esker, N. McRoberts & A. Nelson, "The global burden of pathogens and pests on major food crops", Nature Ecology & Evolution **3** (2019) 430. https://doi.org/10.1038/s41559-018-0793-y.

[3] K. A. Garrett, S. P. Dendy, E. E. Frank, M. N. Rouse & S. E. Travers, "Climate change effects on plant disease: genomes to ecosystems", Annual Review of Phytopathology **44** (2006) 489. https://doi.org/10.1146/annurev.phyto.44.070505.143420.

[4] S. Jauhari & K. K. Agrawal, "A comprehensive review on various plant diseases and impact on crop yield and quality", Journal of Information Systems Engineering and Management **10** (2025) 2468. https://doi.org/10.52783/jisem.v10i38s.6866.

[5] M. Chithambarathanu & M. K. Jeyakumar, "Survey on crop pest detection using deep learning and machine learning approaches", Multimedia Tools and Applications **82** (2023) 42277. https://doi.org/10.1007/s11042-023-15221-3.

[6] H. Ghosh, I. S. Rahat, K. Shaik, S. Khasim & M. Yesubabu, "Potato leaf disease recognition and prediction using convolutional neural networks", EAI Endorsed Transactions on Scalable Information Systems **10** (2023) 1. https://doi.org/10.4108/eetsis.3937.

[7] S. K. Upadhyay & R. Prasad, "Efficient-ViT B0Net: A high-performance lightweight transformer for rice leaf disease recognition and classification", Journal of the Nigerian Society of Physical Sciences **7** (2025) 1. https://doi.org/10.46481/jnsps.2025.2940.

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

[9] L. Weng, Z. Tang, M. F. Sardar, Y. Yu, K. Ai, S. Liang, J. Alhahtani & D. Lyv, "Unveiling the frontiers of potato disease research through bibliometric analysis", Frontiers in Microbiology **15** (2024) 1. https://doi.org/10.3389/fmicb.2024.1430066.

[10] J. H. Sinamenye, A. Chatterjee & R. Shrestha, "Potato plant disease detection: leveraging hybrid deep learning models", BMC Plant Biology **25** (2025) 647. https://doi.org/10.1186/s12870-025-06679-4.

[11] A. Ait Nasser & M. A. Akhloufi, "A hybrid deep learning architecture for apple foliar disease detection", Computers **13** (2024) 116. https://doi.org/10.3390/computers13050116.

[12] S. F. Santoso, S. Hadi, B. Nugroho & I. G. S. Mas Diyasa, "Implementation of hybrid EfficientNet V2 and Vision Transformer for apple leaf diseases classification", Information Technology International Journal **3** (2025) 1. https://doi.org/10.33005/itij.v3i1.42.

[13] C. Zhou, X. Ge, Y. Chang, M. Wang, Z. Shi, M. Ji, T. Wu & C. Lv, "A multimodal parallel transformer framework for apple disease detection and severity classification with lightweight optimization", Agronomy **15** (2025) 1246. https://doi.org/10.3390/agronomy15051246.

[14] I. Pacal, I. Kunduracioglu, M. H. Alma, M. Deveci, S. Kadry, J. Nedoma, V. Slany & R. Martinek, "A systematic review of deep learning techniques for plant diseases", Artificial Intelligence Review **57** (2024) 304. https://doi.org/10.1007/s10462-024-10944-7.

[15] V. Tiwari, R. C. Joshi & M. K. Dutta, "Dense convolutional neural networks based multiclass plant disease detection and classification using leaf images", Ecological Informatics **63** (2021) 101289. https://doi.org/10.1016/j.ecoinf.2021.101289.

[16] M. Xu, J. E. Park, J. Lee, J. Yang & S. Yoon, "Plant disease recognition datasets in the age of deep learning: challenges and opportunities", Frontiers in Plant Science **15** (2024) 1. https://doi.org/10.3389/fpls.2024.1452551.

[17] A. Upadhyay, N. S. Chandel, K. P. Singh, S. K. Chakraborty, B. M. Nandede, M. Kumar, A. Subeesh, K. Upendar, A. Salem & A. Elbeltagi, "Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture", Artificial Intelligence Review **58** (2025) 92. https://doi.org/10.1007/s10462-024-11100-x.

[18] S. R. Trivedi & N. Sharma, "A dynamic deep learning framework for real-time multi-plant, multi-disease detection under diverse environmental conditions", International Journal of Information Technology (2025). https://doi.org/10.1007/s41870-025-02969-0.

[19] Y. Haruna, S. Qin, A. H. Adama Chukkol, A. A. Yusuf, I. Bello & A. Lawan, "Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey", Engineering Applications of Artificial Intelligence **144** (2025) 110057. https://doi.org/10.1016/j.engappai.2025.110057.

[20] Y. N. Kuan, K. M. Goh & L. L. Lim, "Systematic review on machine learning and computer vision in precision agriculture: Applications, trends, and emerging techniques", Engineering Applications of Artificial Intelligence **148** (2025) 110401. https://doi.org/10.1016/j.engappai.2025.110401.

[21] S. A. Salihu, S. O. Adebayo, O. C. Abikoye, F. E. Usman-Hamza, M. A. Mabayoje, B. Brahma & A. Bandyopadhyay, "Detection and classification of potato leaves diseases using convolutional neural network and Adam optimizer", Procedia Computer Science **258** (2025) 2. https://doi.org/10.1016/j.procs.2025.04.159.

[22] G. E. Meyer & J. C. Neto, "Verification of color vegetation indices for automated crop imaging applications", Computers and Electronics in Agriculture **63** (2008) 282. https://doi.org/10.1016/j.compag.2008.03.009.

[23] I. T. Jolliffe & J. Cadima, "Principal component analysis: A review and recent developments", Philosophical Transactions of the Royal Society A **374** (2016) 20150202. https://doi.org/10.1098/rsta.2015.0202.

[24] Y. LeCun, L. Bottou, Y. Bengio & P. Haffner, "Gradient-based learning applied to document recognition", Proceedings of the IEEE **86** (1998) 2278. https://doi.org/10.1109/5.726791.

[25] K. He, X. Zhang, S. Ren, & J. Sun, "Deep Residual Learning for Image Recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90.

[26] M. Tan & Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks", Proceedings of Machine Learning Research **97** (2019) 6105. https://doi.org/10.48550/arXiv.1905.11946.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit & N. Houlsby, "An image is worth 16×16 words: transformers for image recognition at scale", Proceedings of ICLR (2021). https://doi.org/10.48550/arXiv.2010.11929.