SEMARAK ILMU
PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

SEMARAK ILMU
Journal of
ADVANCED RESEARCH IN APPLIED SCIENCES AND ENGINEERING TECHNOLOGY

Enhancing the Resilience of Deep Learning Models for Agricultural Applications: A Study on Adversarial Robustness

Tahajib Jakir Khan[1,*], Prosenjit Chandra Biswas[1], Jannatul Ferdous Momo[1], Rafiul Islam[1], Sazzadur Rahman[2], Shah Md. Tanvir Siddiquee[1]

[1] Department of Computer Science and Engineering, Faculty of Science and Information Technology, Daffodil International University, Dhaka 1216, Bangladesh
[2] Department of Computer Science and Engineering, Faculty of Science, University of Dhaka, Dhaka 1000, Bangladesh

ARTICLE INFO

Article history:
Received 31 December 2025
Received in revised form 8 January 2026
Accepted 14 January 2026
Available online 15 January 2026

Keywords:
CNN-LSTM Hybrid Model; Adversarial Robustness; Betel Leaf Disease Classification; PGD Adversarial Attack; Deep Learning in Agriculture

ABSTRACT

Early and accurate detection of plant diseases is crucial for maintaining crop health and ensuring agricultural productivity. This study investigates the classification of betel leaf and vine diseases using a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) architecture, designed to capture spatial patterns and model structured spatial dependencies by treating CNN-extracted feature representations as ordered sequences within static leaf and vine images, rather than temporal dynamics. To address vulnerabilities to adversarial perturbations, which can mislead standard deep learning models even with visually imperceptible changes, this work incorporates adversarial training based on the Projected Gradient Descent (PGD) method. A curated dataset of betel leaf images, including both natural and adversarially perturbed samples, is used to train and evaluate the model. Experimental results demonstrate that the adversarially trained CNN-LSTM maintains high classification accuracy 96.96%, achieves F1-score of 96.25%, and robustness accuracy under PGD attacks of 92.19%. Class-wise analyses using precision, recall, and F1-score confirm balanced performance across all disease categories, highlighting the model's reliability under challenging input conditions. These findings underscore the importance of integrating robustness-focused strategies in deep learning systems for plant disease detection and provide insights that can guide the development of more robust AI solutions for agricultural imaging applications, while future work is required to assess performance under field conditions and deployment constraints.

1. Introduction

The use of artificial intelligence (AI) in agriculture has gained attention for automated plant disease detection, where early and accurate diagnosis reduces crop loss, improves yield, and supports sustainable practices [1,2]. Betel leaf cultivation, economically important in many regions, is highly susceptible to diseases, highlighting the need for reliable diagnostic systems. Deep learning, particularly convolutional neural networks (CNNs), performs well in image-based plant disease classification [3] but is vulnerable to adversarial perturbations, small, visually imperceptible changes

[*] *Corresponding author.*
E-mail address: Jakir15-14503@diu.edu.bd

https://doi.org/10.37934/araset.55.1.292311

that can cause confident misclassifications [4,5]. To address this, we propose a robust framework integrating CNNs with long short-term memory (LSTM) layers [6] and adversarial training [7]. CNN layers extract discriminative spatial features, while LSTM layers model dependencies to capture disease-related patterns effectively [8]. Iterative gradient-based adversarial samples are incorporated during training to improve robustness and maintain classification performance. Evaluation on a curated betel leaf dataset demonstrates high accuracy and enhanced resistance to adversarial interference, providing a foundation for robust plant disease detection applicable to other crops [9].

To ensure real-world applicability, a betel leaf and vine image dataset was curated from active farms, capturing variations in disease severity, leaf maturity, lighting, and background [10]. Images were classified into Healthy Betel Vine, Healthy Leaf, Rot Disease, and Spot Disease, with data augmentation (rotation, flipping) applied to improve variability and reduce overfitting [11]. An adversarial dataset was generated using projected gradient descent (PGD) [7] and incorporated via a custom data generator alternating between natural and adversarial batches [12]. This dual dataset enabled rigorous evaluation of both accuracy and resilience [13], critical for AI systems in agriculture where errors can cause economic loss [14]. While adversarial robustness is well-studied in computer vision [15], it remains underexplored in agricultural imaging, with unique dataset characteristics challenging model stability and generalization [16].

This study is guided by two research questions: RQ1 examines why conventional CNN-based plant disease models are vulnerable to adversarial perturbations and how this limits their deployment under variable field conditions, while RQ2 investigates whether adversarial training enhances the robustness and performance of a hybrid CNN-LSTM model for betel leaf disease classification. Figure 1 illustrates the study workflow, from motivation and related work to methodology, experiments, evaluation, and conclusions. To support realistic evaluation, a dedicated betel leaf image dataset was collected from active farms, capturing diverse health and disease conditions and augmented for generalization. An adversarial version was generated via projected gradient descent (PGD) and incorporated using a custom data generator alternating between clean and adversarial batches, creating a rigorous dual-mode training environment reflective of practical agricultural challenges.

Despite deep learning's success in plant disease detection, a key gap remains in robustness to adversarial perturbations in agricultural imaging. Conventional CNNs achieve high accuracy on clean data [28] but are highly vulnerable to small adversarial changes [29-31], causing confident misclassifications and reduced reliability under variable field conditions. Although adversarial robustness is widely studied in computer vision [15], its use in crop-specific detection, especially for economically important crops like betel leaf, remains limited. This gap matters because unreliable AI diagnostics can cause misjudgement, delayed intervention, economic loss, and reduced trust among farmers and policymakers [19,20]. This study proposes a hybrid CNN-LSTM model trained on both natural and adversarial images [17,18] to assess adversarial robustness using a custom dataset, apply tailored adversarial training strategies [21,22], and compare performance under clean and perturbed conditions [23-25]. The goal is a robust, deployable architecture that maintains accuracy and stability in real agricultural environments [26,27].

2. Literature Review

Deep learning is promising for plant disease detection, but variable lighting, background noise, and temporal changes limit real-world reliability. CNNs, adversarial vulnerabilities, and CNN-LSTM hybrids have been studied, yet robustness in agricultural settings remains underexplored. This study evaluates whether adversarially trained CNN-LSTM models sustain accuracy and stability under realistic perturbations.

2.1 Deep Learning in Agricultural Applications

The application of deep learning in agriculture has grown rapidly over the past decade, driven by the need for scalable, accurate, and automated solutions in crop monitoring and disease detection. Kamilaris and Prenafeta-Boldú [1] highlighted the broad potential of deep learning in agriculture, emphasizing its role in tasks such as yield prediction, land use classification, and plant disease identification. Similarly, Liakos et al. [2] reviewed how machine learning supports decision-making across the agricultural value chain, from field to market.

Plant disease detection, in particular, has received considerable attention. Mohanty et al. [3] demonstrated the effectiveness of convolutional neural networks (CNNs) in identifying plant diseases from leaf images. Their findings revealed that CNNs could achieve high classification accuracy, making them suitable for practical field applications. Ferentinos [14] further supported this by applying deep learning to real-world datasets and achieving robust performance under varied conditions.

Despite these promising results, most studies assume clean input conditions, limiting their applicability in complex environments. Barbedo [10] stressed that real-world agricultural images often contain noise, inconsistencies in lighting, and background clutter, which can degrade model performance. This creates a gap between laboratory accuracy and field reliability.

2.2 Adversarial Robustness and Its Relevance to Agriculture

While deep learning models excel in image classification, their vulnerability to adversarial examples has raised significant concerns. Yuan, Xiaoyong, et al. [4] were among the first to reveal that small, often imperceptible, perturbations in input images can mislead neural networks. Madry et al. [7] introduced adversarial training as an effective defense mechanism, where models are trained on adversarial samples to improve robustness. Croce and Hein [13] further demonstrated that many proposed defenses are less effective under strong evaluation attacks, reinforcing the importance of building inherently robust architectures.

Although adversarial robustness has been extensively studied in domains like cybersecurity and autonomous driving, its application in agriculture remains relatively underexplored. Singh et al. [9] observed that robust model behavior is critical for plant stress detection, especially when environmental variability is high. Our research addresses this gap by applying and evaluating adversarial training within agricultural disease detection.

2.3 Hybrid Deep Learning Architectures for Temporal and Spatial Features

Traditional CNN models are effective at extracting spatial features, but they often lack the ability to capture sequential dependencies or dynamic environmental changes. To overcome this, hybrid models that combine CNNs with recurrent neural networks such as LSTMs have been proposed. Xingjian et al. [6] introduced the ConvLSTM model, which merges convolutional operations with LSTM cells to process spatiotemporal data effectively. Zhu et al. [8] highlighted the applicability of such models in remote sensing, emphasizing their capacity to handle sequential visual patterns.

Although CNN-LSTM architectures have been successfully applied in domains such as remote sensing and spatiotemporal modeling, their adoption in plant disease classification remains limited, and their integration with adversarial training in agricultural imaging has not been systematically explored. Additionally, real-world agricultural datasets often exhibit limited diversity and class imbalance, which can affect model training and generalization.

3. Methodology

This section outlines the methodology used to investigate the adversarial vulnerability of CNN-based models and the effectiveness of adversarial training in improving the robustness of a CNN-LSTM architecture for classifying betel leaf diseases. The methodology is structured into five components: problem formulation, data preparation, model design, adversarial attack generation, and adversarial training and evaluation.

Figure 1 synthesizes the complete methodological pipeline of this study, illustrating how raw betel leaf images are systematically transformed into robust disease predictions through tightly integrated stages of preprocessing, model development, and adversarial learning. Figure 1 highlights the dual training strategy, clean and PGD-based adversarial learning, which reinforces the CNN-LSTM architecture to extract resilient spatial features and model structured dependencies across spatial feature sequences, while exhibiting improved robustness under adversarial perturbations. Overall, this end-to-end workflow demonstrates a coherent and practical approach to addressing adversarial vulnerability in agricultural image classification, directly aligning the methodological design with the study's robustness-driven objectives.

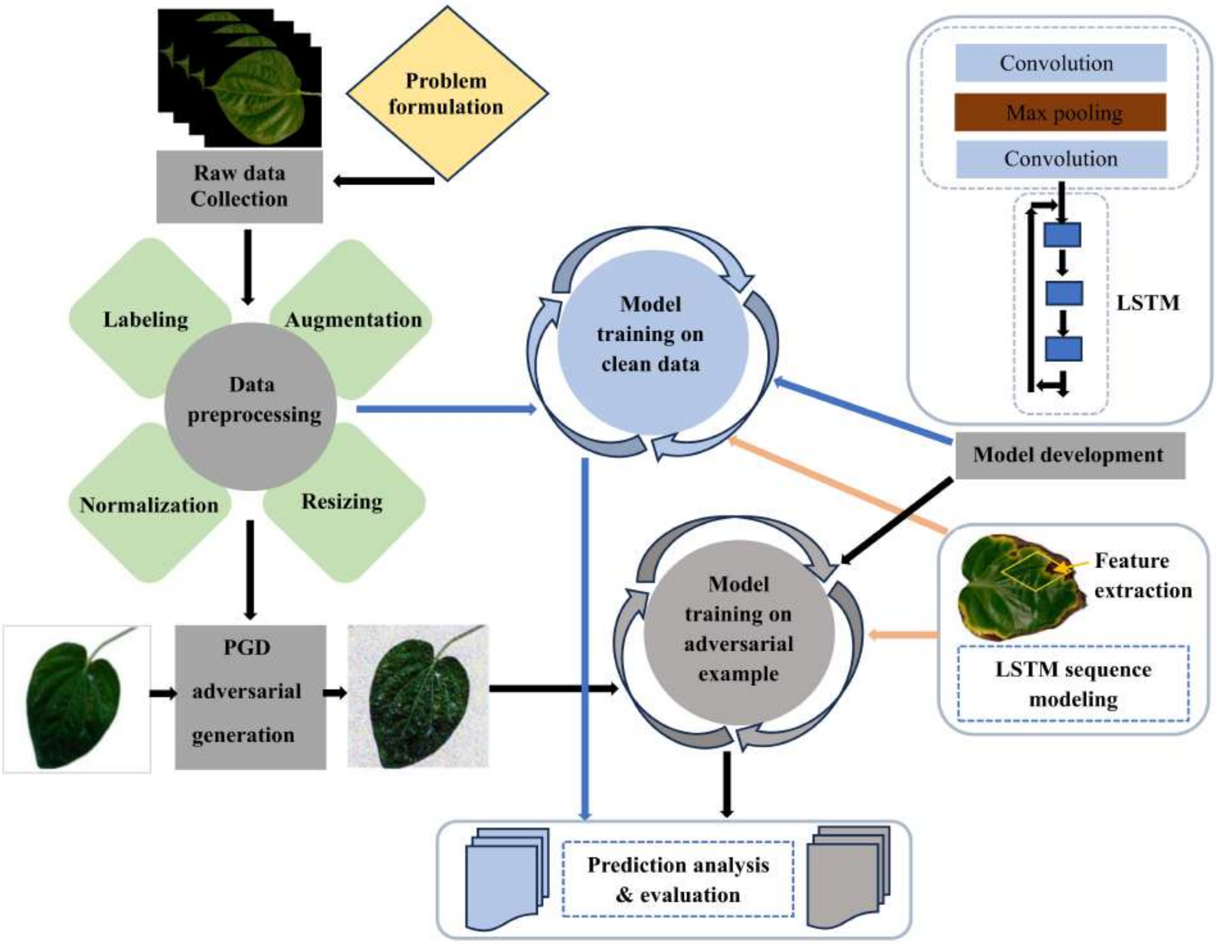


Fig. 1. Overview of the proposed CNN-LSTM framework with PGD-based adversarial training

3.1 Problem Formulation

This study aims to address two primary questions: (1) how standard CNN models respond when input images are subtly altered by adversarial perturbations (RQ1), and (2) whether incorporating adversarial training can improve the stability and reliability of a CNN-LSTM model for classifying betel leaf diseases (RQ2). The task is formulated as a supervised image classification problem, where the model f_θ maps an input image $x \epsilon \mathbb{R}^{H \times W \times 3}$ to a predicted label $y \in \{1,2,\ldots,C\}$. Model parameters θ are optimized by minimizing the categorical cross-entropy loss.

$$\mathcal{L}(x,y) = -\log f_\theta(x)_y \tag{1}$$

In this context, adversarial robustness refers to the ability of the classification model to preserve correct disease predictions when input images are perturbed within a bounded and visually imperceptible range, denoted x_{adv}, constrained by a small perturbation limit $\| x_{\text{adv}} - x \|_\infty \leq \epsilon$. Such perturbations can expose vulnerabilities in the model's decision boundaries, which, if unaddressed, may lead to incorrect classifications in real-world agricultural applications.

The problem formulation establishes a framework for analysing the influence of adversarial inputs on model performance and for exploring whether augmenting the training set with adversarial examples can enhance the model's resilience. This conceptual foundation informs the subsequent methodological steps, including data preparation, model architecture design, adversarial example generation, and robustness-focused training, providing a systematic approach to answer both RQ1 and RQ2.

3.2 Data Collection and Preprocessing

The dataset used in this study comprises images of betel leaves and their supporting stems, organized into four functional categories: healthy leaf, healthy betel vine, spot disease, and rot disease. While the primary focus is on leaf health, images of vines are included alongside healthy betel leaves to provide additional structural context. The visual patterns of vines help the model distinguish between healthy growth and early signs of disease, as vines often reflect subtle cues associated with leaf pathology. By incorporating both leaf and vine features, the model is better equipped to learn the anatomical relationships and contextual information necessary for accurate disease classification, enhancing its performance under real-world farm conditions where leaves and vines are captured together. Figure 2 presents representative samples from each dataset class, illustrating variations in leaf texture, vein structure, color intensity, and visible disease symptoms. The samples reflect the inherent heterogeneity of real-world agricultural imagery, including differences in illumination, viewing angles, and background conditions. Such diversity is characteristic of images captured in natural farm environments and underscores the need for robust feature learning in plant disease classification. To standardize these visually diverse images, we implemented a multi-stage preprocessing pipeline. First, all images were resized to 224 × 224 pixels using bilinear interpolation. This resolution balances detail preservation with computational feasibility [33], while ensuring compatibility with CNNs that expect fixed-dimension inputs. Next, pixel intensities were normalized to the [0, 1] range, providing numerical stability and reducing sensitivity to variable lighting conditions. This normalization is achieved through the standard transforms.ToTensor() routine, which ensures consistent pixel scaling across training, validation, and testing sets.

To increase generalization capability, we applied controlled augmentation operations, random rotations, horizontal flips, and modest brightness adjustments. Other works have explored generative approaches such as GANs for creating additional synthetic samples to enrich training datasets [24]. These transformations simulate common field-level variations such as angled captures, sunlight fluctuations, and natural movements of leaves due to wind. This step effectively broadens the data distribution and encourages the model to learn more stable representations instead of memorizing class-specific patterns.

Collectively, the preprocessing strategy ensures that the dataset is clean, consistent, and sufficiently diverse to support the training of robust deep learning models. The visual variety highlighted in Figure 2 underscores the necessity of such preprocessing measures, as it directly influences downstream model stability and resilience against adversarial manipulations introduced in later sections.

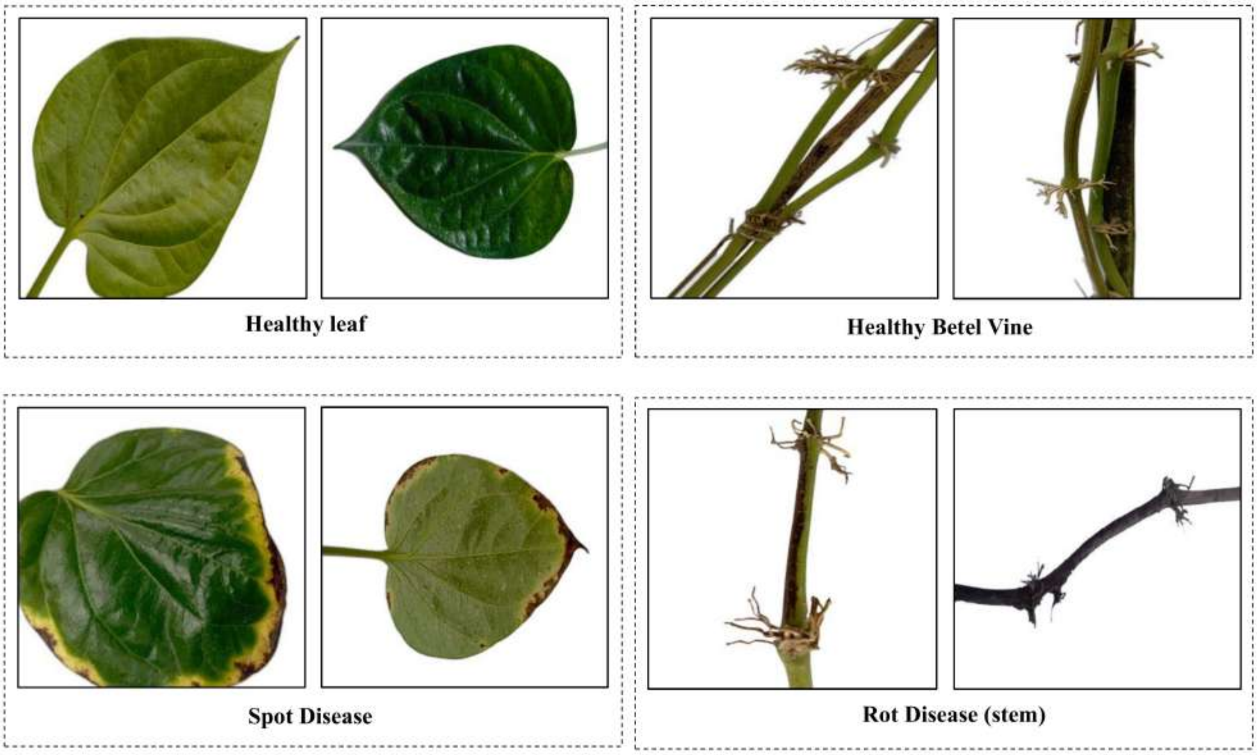


Fig. 2. Representative samples from the betel leaf and vine disease dataset

3.3 Model Architecture and Training

The proposed architecture integrates convolutional feature extraction with LSTM-based dependency modeling to capture both local spatial patterns and structured relationships among high-level spatial features. Here, the LSTM operates on ordered spatial feature sequences rather than temporal data.

Figure 3 outlines the full pipeline, showing the progression of transformations from raw images to final class probabilities. It presents a schematic overview of the architecture, highlighting the progression from convolutional feature extraction to sequential dependency modelling and final classification. This model detail configuration analysed in feature extraction module, sequential modelling and classification layer sections.

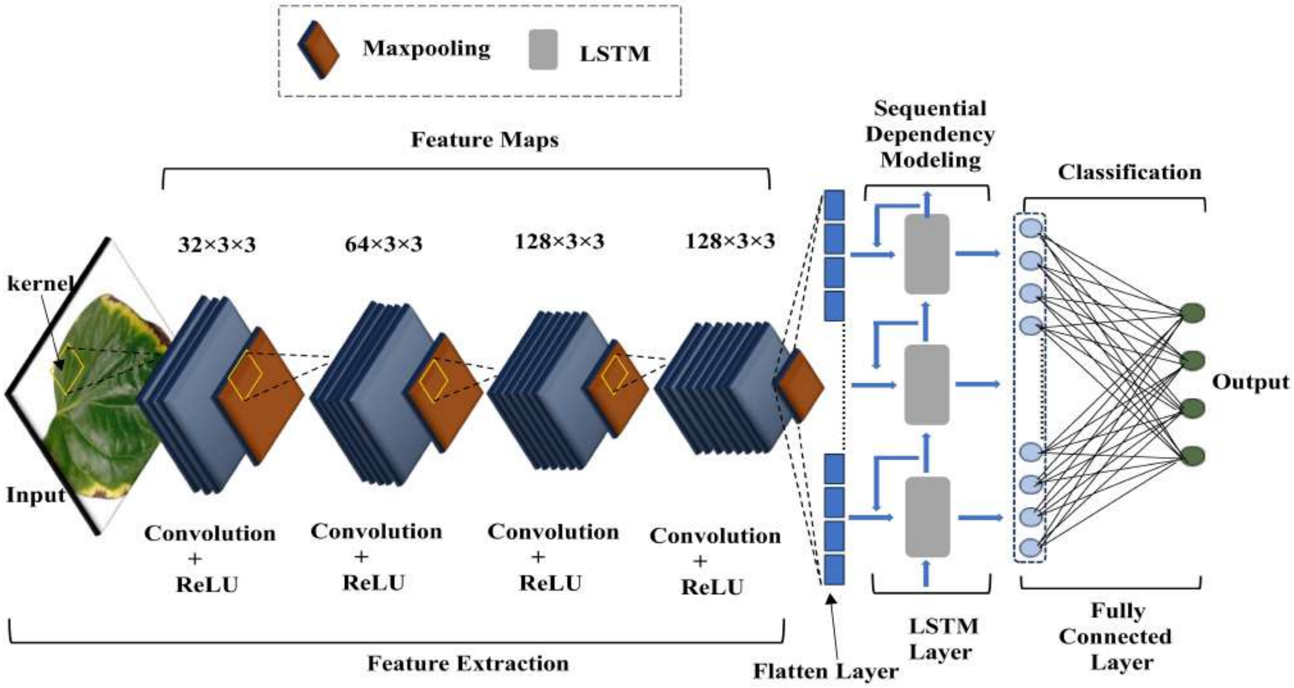


Fig. 3. Architecture of the proposed CNN-LSTM model for betel leaf disease classification

3.3.1 CNN feature extraction module

Given an input RGB image,

$$x \in R^{224\times224\times3}, \tag{2}$$

The CNN module transforms it into a compact feature tensor through a series of convolution–normalization-activation-pooling blocks.

For each convolutional block i, the operation is:

$$F_i = MaxPool(\sigma\left(BN\left(Conv_i(F_{i-1})\right)\right)) \tag{3}$$

Where, is a convolution with kernel size 3×3, BN is batch normalization, $\sigma(\cdot)$ denotes the ReLU activation, $F_0 = x$. The filter depths are 32, 64, 128, and 128 respectively. Each convolution increases the representational capacity of the network, allowing earlier layers to capture coarse edges and colour transitions, while deeper layers respond to lesion boundaries, necrotic textures, mould -like spots, and structural vine characteristics.

Residual shortcuts (shown in Figure 3) connect selected convolutional blocks and are defined as:

$$F_i^{res} = F_i + W_i F_{i-1}, \tag{4}$$

where W_i matches the dimensions via a 1×1convolution. This improves gradient flow and stabilizes deeper feature learning important given the heterogeneous nature of the dataset (as depicted in Figure 2) [34,35].

After the final block, the feature tensor has shape:

$$F_{CNN} \in R^{H'\times W'\times128} \tag{5}$$

with $H' = W' = 14$after sequential pooling.

To provide an intuitive understanding of how the convolutional layers progressively encode visual information, Figure 4 visualizes representative feature maps extracted at different depths of the CNN. These feature activations illustrate the hierarchical nature of the learned representations, moving from low-level appearance cues toward more abstract and semantically meaningful patterns relevant to disease characterization.

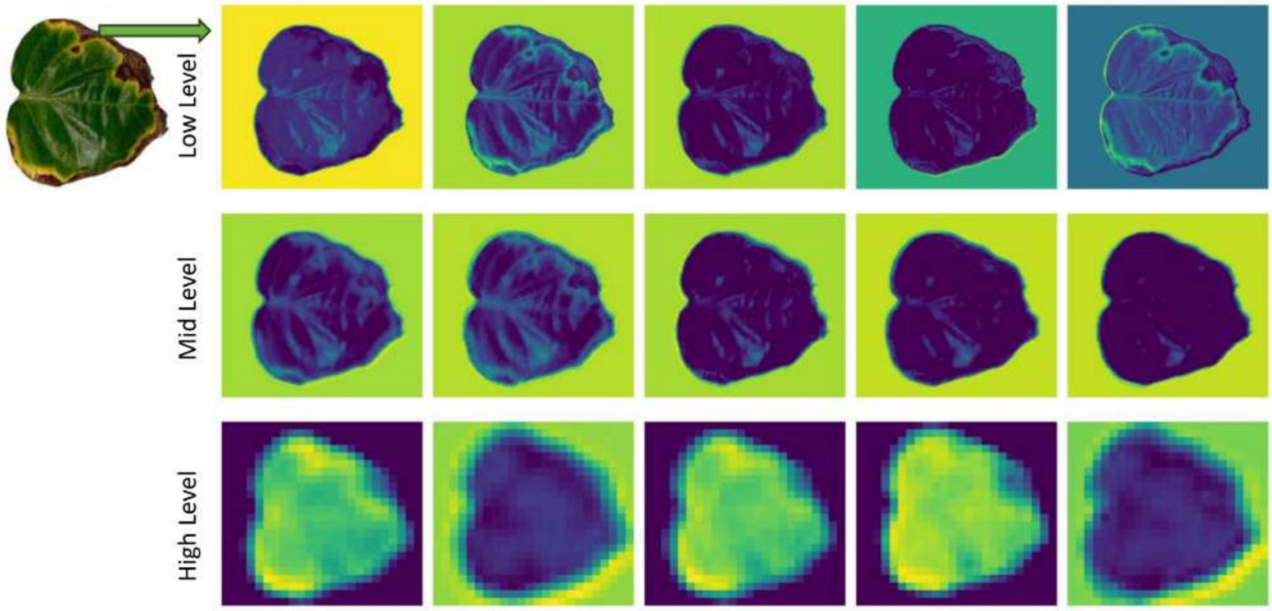


Fig. 4. Visualization of hierarchical feature representations extracted by the CNN

As illustrated in Figure 4, the early convolutional layers primarily respond to low-level visual attributes such as edges, colour gradients, and basic vein structures, while intermediate layers begin to emphasize localized texture irregularities and shape distortions. In contrast, deeper layers exhibit increasingly coarse but semantically focused activations, highlighting regions associated with disease symptoms such as lesion concentration, tissue degradation, and abnormal pigmentation. This progressive abstraction demonstrates the CNN's ability to filter out irrelevant background information while preserving diagnostically salient regions. The incorporation of residual connections further stabilizes this hierarchical learning process by maintaining feature continuity across layers, which is particularly important given the variability in illumination, texture, and structural composition present in the dataset. Collectively, these characteristics ensure that the extracted feature tensor provides a robust and information-rich representation, well suited for subsequent sequential modelling in the LSTM module.

3.3.2 Reshaping for sequential modelling and LSTM-based dependency modelling

To exploit spatial continuity across the leaf surface, the CNN output is reshaped into a sequence:

$$S = Reshape(F_{CNN}) \in R^{T\times d}, \quad (6)$$

Where T=14×14 = 196 sequence steps, d= 128 feature per step.

This converts the 2D spatial map into a 1D ordered sequence. The reshaping follows a consistent raster-scan ordering (row-major), ensuring reproducibility [36-37]. This transformation allows the LSTM to learn structured spatial relationships among neighbouring regions of the leaf surface, such as correlated lesion patterns and texture transitions, without assuming temporal evolution. While LSTM networks are conventionally used for temporal sequence modelling, in this work they are adapted to model ordered spatial dependencies among high-level feature vectors extracted by the CNN backbone. By reshaping CNN feature maps into sequential representations, the LSTM captures long-range structural patterns related to disease morphology, vein structure, and texture continuity, which are difficult to model using convolutional layers alone.

The sequence S is fed into a two-layer LSTM with hidden dimension h. For each time step t:

$$h_t, c_t = LSTM(s_t, h_{t-1}, c_{t-1}), \tag{7}$$

Where h_t is the hidden state, c_t is the cell state.

The LSTM gates are:

$$f_t = \sigma(W_f s_t + U_f h_{t-1} + b_f, \tag{8}$$

$$i_t = \sigma(W_i s_t + U_i h_{t-1} + b_i, \tag{9}$$

$$\tilde{c}_t = tanh(W_c s_t + U_c h_{t-1} + b_c, \tag{10}$$

$$o_t = \sigma(W_o s_t + U_o h_{t-1} + b_o, \tag{11}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{12}$$

$$h_t = o_t \odot tanh(c_t)\,, \tag{13}$$

which together control how relevant spatial features propagate across the leaf surface. The final hidden state h_T serves as a compact representation of all contextual dependencies (e.g., co-occurring lesion regions, rot-affected areas, and disrupted vein structures).

To further illustrate how the LSTM processes the spatial feature sequence derived from the CNN output, a visualization of LSTM activations across sequence steps and feature dimensions is presented. This representation provides a qualitative illustration of how the LSTM aggregates spatial feature information across ordered feature sequences corresponding to different spatial locations on the leaf surface, highlighting variations in activation intensity as these dependencies are integrated. By examining these activation patterns, it becomes possible to qualitatively assess how the LSTM encodes contextual information beyond isolated local features.

The activation structure in Figure 5 shows that the LSTM forms a structured activation topology across spatial sequence steps and feature dimensions, indicating that spatial dependencies are hierarchically prioritized rather than uniformly spread across the leaf surface. Certain sequence positions produce stronger responses, meaning the model selectively emphasizes diagnostically salient regions, such as lesion clusters, discoloration, or vein disruptions. Instead of treating each spatial location independently, the LSTM integrates information across neighbouring regions, allowing localized anomalies to influence the global representation. This is valuable for agricultural disease classification, where symptoms appear as spatially correlated patterns rather than isolated pixels. The final hidden state thus encodes a holistic spatial context, capturing interactions between healthy and diseased regions and complementing convolutional feature extraction. By converting spatial feature maps into a dependency-aware representation, the LSTM enhances generalization under challenging visual conditions and supports a more robust classification stage.

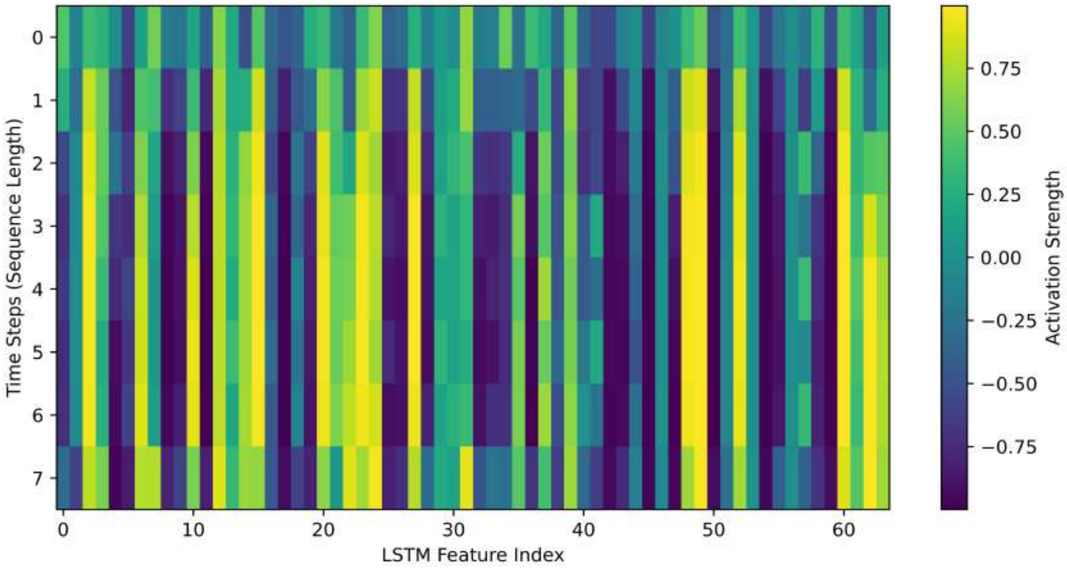


Fig. 5. LSTM-based sequential aggregation of spatial feature representations

3.3.3 Classification layer and training configuration

The classifier maps the final state to class probabilities:

$$\hat{y} = Softmax(Wh_T + b), \tag{14}$$

Where,

$$\hat{y}k = \frac{exp\ (z_k)}{\sum_{j=1}^{4} exp\ (z_j)}\ , k \in \{1,2,3,4\}. \tag{15}$$

The model is trained to minimize the standard cross-entropy:

$$\mathcal{L}(x, y) = -\log(\tilde{y}_y)\ . \tag{16}$$

Here, $\tilde{y}_y$ denotes the predicted probability corresponding to the true class y. The softmax function converts the network's outputs into a probability distribution, while the categorical cross-entropy loss quantifies the discrepancy between predicted and true labels. Minimizing this loss guides the model to assign higher confidence to the correct class, effectively translating the extracted spatial and sequential features into accurate predictions. The subsequent table summarizes the training configuration that supports this optimization process.

The table (Table 1) summarizes the main components of the training setup, reflecting the deliberate design choices for both convergence efficiency and model robustness. Beyond the listed parameters, it is important to note that the alternating batch strategy between clean and adversarial images enables the model to simultaneously capture natural variations and defend against perturbations, promoting a balanced feature representation. Additionally, the selected learning rate schedule provides a gradual refinement of weight updates, reducing the risk of overshooting during optimization and ensuring smoother convergence trajectories. Collectively, these training considerations form a foundation for achieving high classification performance while maintaining resilience to adversarial disturbances.

Table 1
Training configuration and optimization settings of the proposed CNN-LSTM model

Parameter	Values	Purpose / Description
Optimizer	Adam	Provides adaptive learning rates for stable and efficient convergence.
Initial Learning Rate	0.0005	Controls the initial step size during gradient updates.
Learning Rate Schedule	Step decay (×0.9 after epoch 10)	Gradually refines learning to improve convergence.
Batch Size	100	Balances computational efficiency and gradient stability.
Loss Function	Categorical Cross-Entropy	Optimizes multi-class classification performance.
Evaluation Metrics	Accuracy, Precision, Recall, F1-score	Quantitatively evaluates classification performance.
Number of Epochs	20	Ensures sufficient learning without overfitting.
Training Strategy	Alternating clean and adversarial batches	Improves robustness while preserving clean-data accuracy.
Hardware Platform	Google Colab with NVIDIA Tesla T4 GPU	Accelerates training and ensures reproducibility.

3.4 Adversarial Example Generation

Adversarial examples in this study are generated using the Projected Gradient Descent (PGD) procedure, which constructs perturbations through an iterative update process while constraining them within an ℓ∞-bounded region. Let x denote a clean input image and y its corresponding class label. The initialization step sets the adversarial counterpart to $x_0^{adv} = x$. For each iteration $t = 0,1,\dots,T-1$, the adversarial sample is refined by applying a gradient-ascent step that maximizes the classification loss, followed by a projection that enforces the perturbation limit. The update is defined as:

$$x_{t+1}^{adv} = \Pi_{B_\epsilon(x)}(x_t^{adv} + \alpha \cdot Sign(\nabla_x \mathcal{L}(f(x_t^{adv}), y))) \tag{17}$$

where L represents the cross-entropy loss, $f(\cdot)$ is the CNN-LSTM classification model, and $\nabla_x \mathcal{L}$ denotes the gradient of the loss with respect to the input. The parameter α controls the step size (set to 0.01), while ϵdetermines the upper bound on the allowable perturbation (fixed at 0.02). The projection operator $\Pi_{B_\epsilon(x)}(\cdot)$ ensures that the updated sample remains within the valid ϵ-ball around the clean image. This projection is implemented as:

$$\Pi_{B_\epsilon(x)}(z) = min\ (max\ (z, x-\epsilon), x+\epsilon), \tag{18}$$

followed by clamping the resulting pixel intensities to the valid image domain [0,1]. Together, these steps guarantee that the perturbation remains visually imperceptible while still enabling the iterative refinement needed for a strong adversarial effect. Repeated iterations (in this study, T=80) allow PGD to follow a more accurate ascent path through the input space than single-step attacks such as FGSM, thereby yielding adversarial examples that exhibit greater destructive potential while still respecting the ℓ∞ constraint.

This formulation also clarifies the conceptual relationship between PGD and the Basic Iterative Method (BIM). Although BIM applies iterative FGSM updates with clipping, PGD explicitly formulates this process as a projection onto the ℓ∞-bounded constraint set at every iteration, resulting in a more general and theoretically grounded framework. PGD, in contrast, performs this projection at every step, which prevents the perturbation from drifting outside the allowable region. Thus, BIM can be viewed as a special or restricted case of PGD, whereas the version adopted in this work adheres to the full and more robust PGD framework.

Figure 6 presents adversarial examples generated at increasing values of ϵ, showing that perturbed images remain visually indistinguishable from the originals despite causing incorrect model predictions. This highlights a key vulnerability of deep learning systems, where imperceptible perturbations can induce misclassification. As ϵ increases, adversarial strength grows without introducing visible artifacts. These samples are used directly in the adversarial training process described in Section 3.5 and form a core component of the robustness enhancement pipeline.

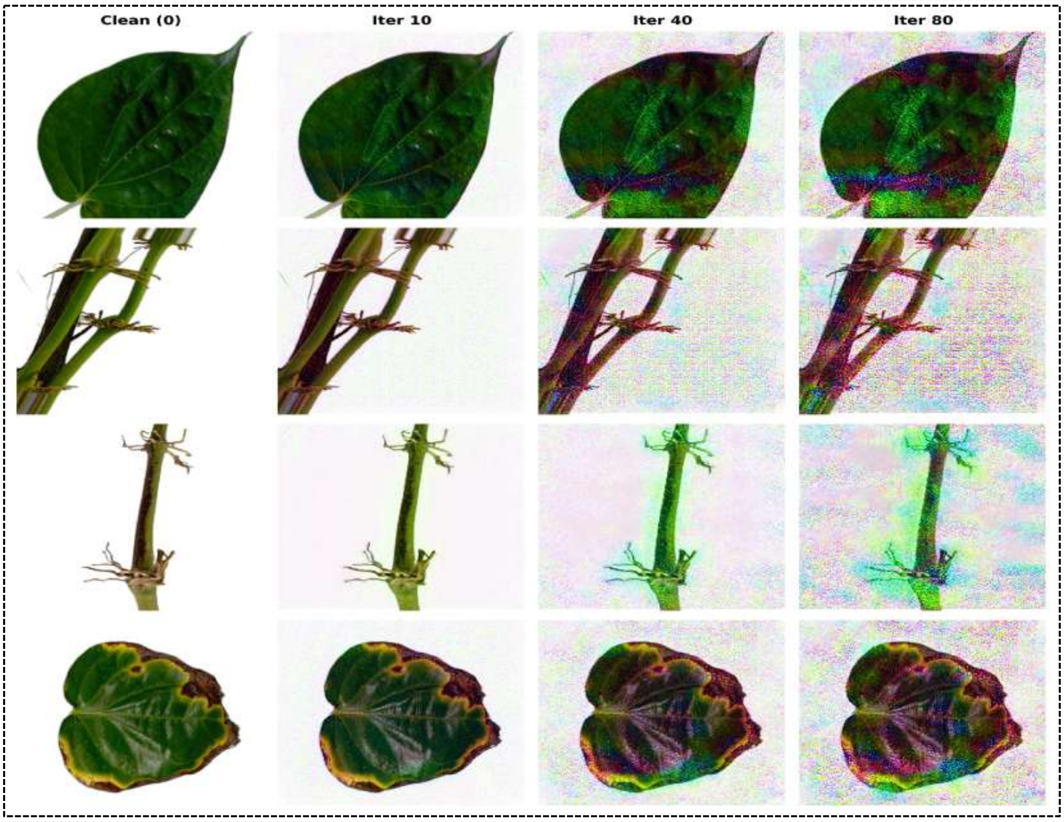


Fig. 6. Visual comparison of original and PGD-generated adversarial images under varying perturbation budgets (ϵ)

3.4 Adversarial Example Generation

To defend against adversarial attacks and improve model robustness (in alignment with RQ2), we incorporate adversarial training into the learning process. At each training step, adversarial examples are generated during training and integrated with clean samples, used alongside clean images to update model weights.

The training loop alternates between clean and perturbed batches, minimizing the combined loss:

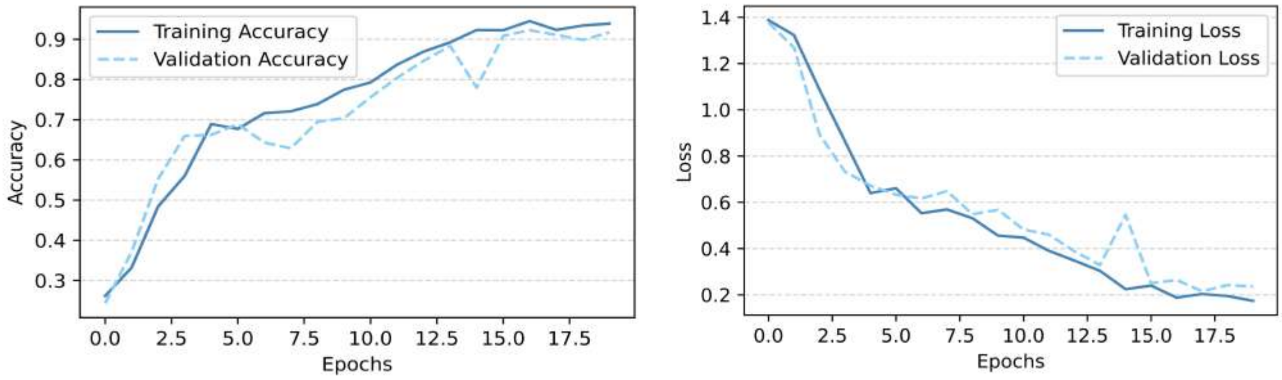
$$\mathcal{L}_{total} = \mathcal{L}(x, y) + \beta.\mathcal{L}(x_{adv}, y) \tag{19}$$

Where, x: clean image, x_adv: adversarial image, β: scalar weighting factor empirically set to balance clean accuracy and adversarial robustness during training.

Model performance is evaluated using accuracy on both clean and adversarial datasets, along with class-wise precision, recall, and F1-scores to capture detailed classification behaviour. Robust

Accuracy is measured as the classification accuracy on adversarial samples generated at specified ϵ levels, providing a direct quantitative indicator of model resilience to adversarial perturbations [38].

The learning curves show that the adversarially trained model retains high accuracy and generalization ability, even under perturbation. Metrics across classes remain balanced, confirming that no specific class dominates or collapses under attack conditions. The figure demonstrates consistent model convergence, with minimal overfitting and stable validation loss. These results indicate that adversarial training not only enhances robustness but also preserves classification fidelity validating the methodology's effectiveness.



(a) (b)

Fig. 7. Training and validation learning curves of the adversarially trained CNN-LSTM model

Overall, the integration of PGD-based adversarial training within the proposed CNN-LSTM architecture leads to improved robustness against iterative adversarial perturbations while maintaining strong performance on clean data. By jointly leveraging convolutional feature extraction for local pattern learning and sequential modelling for spatial dependency aggregation, the model is able to learn representations that are less sensitive to bounded input distortions. The experimental results indicate that exposure to adversarial samples during training contributes to more stable decision boundaries, supporting reliable classification under both natural and adversarial conditions. This balance between accuracy and robustness makes the proposed approach suitable for practical agricultural disease classification scenarios, where input variability and potential perturbations are expected.

4. Results and Discussion

This section presents the evaluation results of the CNN-LSTM model on both clean and adversarially perturbed data. The discussion is structured to address two core research questions: (1) assessing the vulnerability of conventional CNN-based models to adversarial inputs (RQ1), and (2) evaluating whether adversarial training improves the model's robustness and generalization (RQ2).

4.1 Performance Analysis on the Clean Dataset

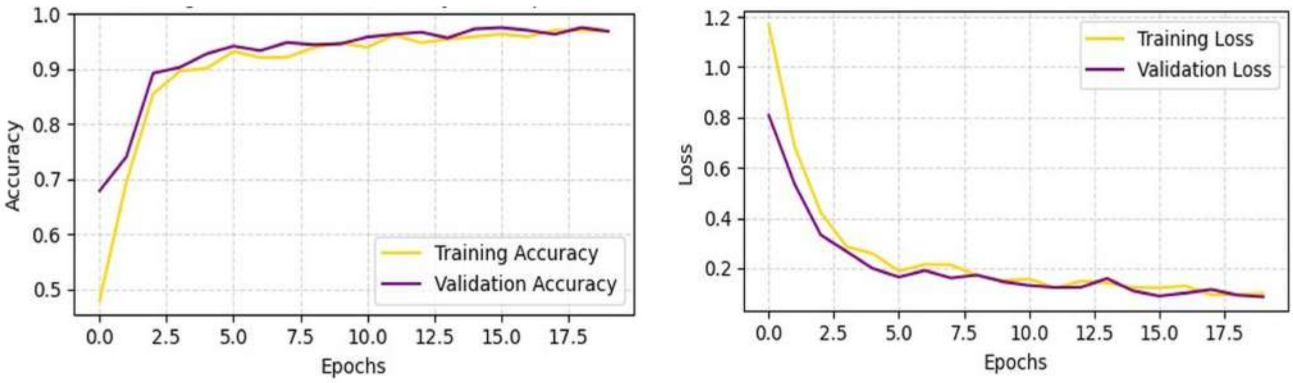
The performance of the proposed CNN-LSTM model is first examined under clean, unperturbed input conditions to establish a reliable baseline for subsequent robustness analysis. The analysis is divided into two parts: first, an examination of the learning dynamics during training and validation, and second, an assessment of the model's generalization capability on an independent test set. Establishing strong and stable performance on clean data is essential, as it provides the baseline reference for analysing adversarial vulnerability and robustness in later sections.

4.1.1 Training and validation performance

To analyse the learning behaviour of the proposed model, training and validation accuracy and loss were monitored across 20 epochs. This evaluation provides insight into convergence characteristics, optimization stability, and the model's ability to generalize beyond the training data.

Figure 8 illustrates the evolution of training and validation accuracy and loss throughout the learning process. At the initial stage of training, the model exhibits relatively low accuracy (approximately 40.8%), which reflects the complexity of the multi-class classification task and the variability inherent in real-world agricultural imagery. However, a rapid improvement is observed within the first few epochs, indicating that the network progressively learns discriminative spatial features and structured spatial dependencies across feature representations.

As training progresses, both training and validation accuracy increase steadily and converge to final values of 96.13% and 96.88%, respectively. The close alignment between these curves suggests effective generalization and an absence of significant overfitting. This observation is reinforced by the corresponding loss trends: the training loss decreases consistently from 1.2942 to 0.1157, while the validation loss declines from 0.8084 to 0.0868 by the final epoch. Importantly, no divergence between training and validation loss is observed, indicating stable optimization despite the hybrid model's depth and complexity. Minor fluctuations in validation accuracy during intermediate epochs are expected given the heterogeneous nature of field-acquired images, which include variations in illumination, background clutter, and subtle inter-class similarities. Overall, the training dynamics demonstrate that the CNN-LSTM architecture achieves smooth convergence and learns a compact yet discriminative representation under clean input conditions.



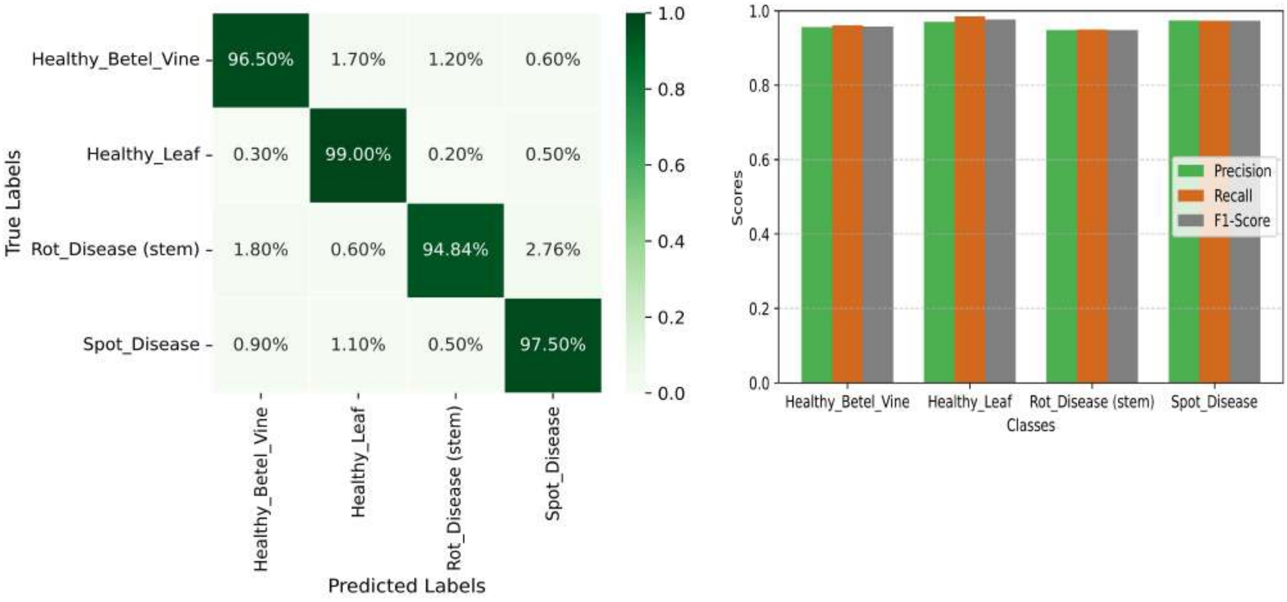
(a) (b)

Fig. 8. Accuracy and loss convergence of the CNN-LSTM model under clean input conditions

4.1.2 Test-time performance on clean data

The evaluation of the CNN-LSTM model on the clean test dataset confirms its strong generalization capability across all classes of betel leaf conditions. The overall test accuracy of 96.96% with a corresponding loss of 0.1056 closely mirrors the validation performance observed during training, indicating that the learned feature representations are robust and not overfitted to the training data. Analysis of the confusion matrix (Figure 9a) reveals that the model maintains a high degree of class-wise fidelity, correctly identifying nearly all instances of healthy leaves, healthy betel vine, rot disease, and spot disease. Misclassifications are minimal, with the largest confusion occurring between rot disease and healthy vine, reflecting the subtle visual similarities in early stages of leaf degeneration, a nuance that the model largely overcomes.

The detailed examination of precision, recall, and F1-score across classes (Figure 9b) further highlights the model's effectiveness. Precision values ranging from 94.8% to 97.4% demonstrate that the model is adept at limiting false positives, while recall values between 94.9% and 98.5% indicate a strong capability to correctly identify true disease cases. The slightly higher recall observed for certain classes suggests a conservative tendency, favouring the detection of actual disease instances over the misclassification of healthy samples, a behaviour that aligns with practical agricultural priorities, where failing to detect a disease can have more severe consequences than issuing a false alert. The high and balanced F1-scores, spanning 94.8% to 97.7%, reflect that the model maintains an equitable trade-off between precision and recall across all categories, without disproportionately favouring any single class.



(a) *(b)*

Fig. 9. Test-phase predictive performance of the CNN-LSTM model under clean conditions

Collectively, these observations provide compelling evidence that the CNN-LSTM architecture successfully captures both local and global spatial dependencies within the betel leaf imagery, resulting in a model that performs consistently under ideal, unperturbed conditions. This establishes a reliable baseline for evaluating adversarial robustness, addressing RQ1 by demonstrating that, in the absence of perturbations, the model achieves both high accuracy and balanced class-wise performance, effectively supporting precise and trustworthy disease diagnosis in real-world agricultural scenarios.

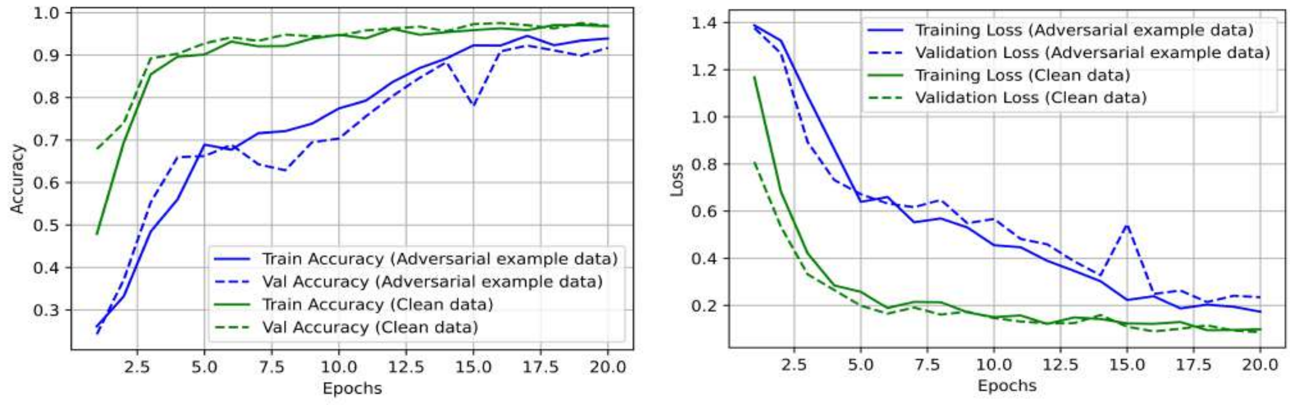
4.2 Adversarial Robustness Evaluation and Performance Under Attack

This section empirically evaluates the behaviour of the proposed CNN-LSTM model under adversarial conditions generated using a PGD-based attack. Model performance is assessed through learning dynamics during adversarial training, test-time behaviour on adversarially perturbed samples, and class-wise robustness metrics. The analysis focuses exclusively on measurable outcomes derived from accuracy, loss, confusion matrices, and class-level performance indicators, without introducing interpretive or causal explanations. This structured evaluation provides a quantitative basis for subsequent discussion of robustness and model behaviour under adversarial perturbations.

4.2.1 Learning dynamics under adversarial training

Figure 10 illustrates the training and validation accuracy and loss trajectories of the CNN-LSTM model under clean and PGD-based adversarial training over 20 epochs. As shown in Figure 10(a), learning under adversarial inputs progresses more gradually than under clean conditions. Training and validation accuracy for adversarial samples increase steadily across epochs, with lower initial values and delayed convergence compared to clean data. In contrast, clean-data accuracy rises sharply within the early epochs and stabilizes at higher levels. Despite the slower improvement under adversarial training, both training and validation accuracy curves exhibit consistent upward trends without abrupt drops or oscillatory behaviour.

The loss curves in Figure 10(b) further highlight the increased optimization difficulty introduced by adversarial examples. Training and validation loss for adversarial data begin at higher magnitudes and decrease more slowly than their clean-data counterparts. Nonetheless, loss reduction remains monotonic overall, and validation loss closely tracks training loss throughout the training process. Minor fluctuations observed in adversarial validation loss do not persist across epochs and do not indicate divergence or instability.



(a) (b)

Fig. 10. Comparative learning trajectories of the CNN-LSTM model under clean and adversarial training

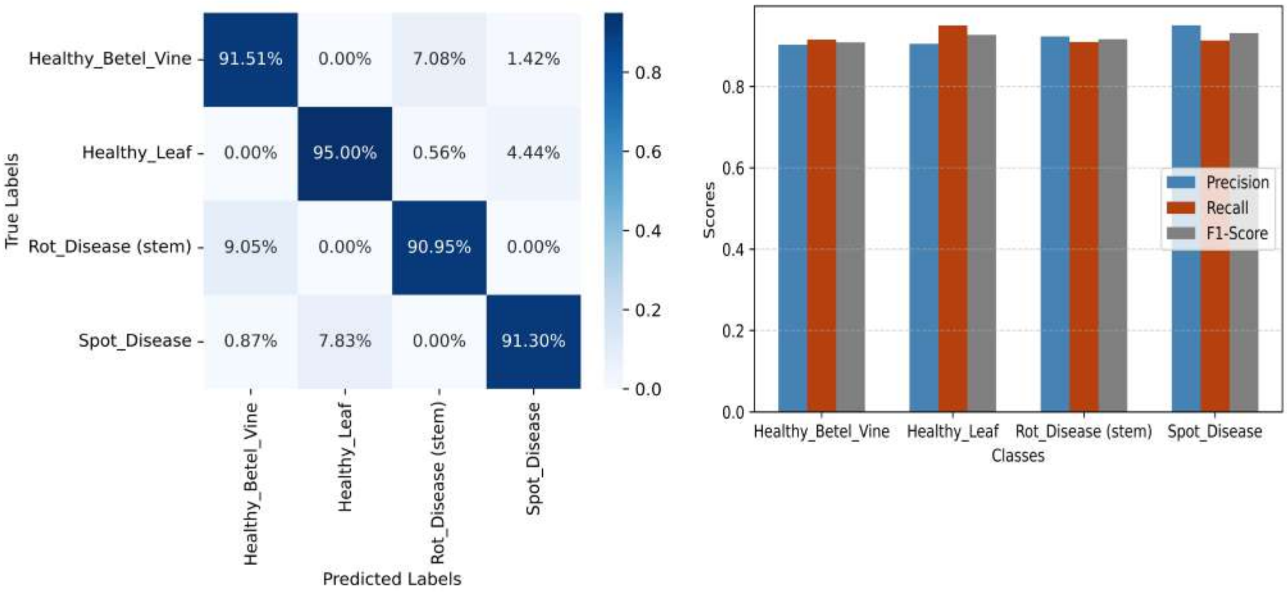
Across both accuracy and loss metrics, no evidence of training collapse, overfitting, or validation instability is observed under adversarial training. The learning dynamics remain controlled and stable, despite the increased complexity of optimizing on perturbed inputs.

4.2.2 Adversarial test performance and class-wise robustness

Figure 11 summarizes the performance of the CNN-LSTM model on the PGD-perturbed test set, combining the confusion matrix (Figure 11a) with class-wise precision, recall, and F1-score metrics (Figure 11b). Together, these results provide a detailed empirical characterization of model behaviour under adversarial evaluation.

The confusion matrix in Figure 11(a) shows that correct predictions remain concentrated along the main diagonal for all four classes, indicating preserved classification capability under adversarial perturbations. Healthy Betel Vine samples achieve a correct classification rate of 91.51%, with the majority of misclassifications directed toward the Rot Disease (stem) class (7.08%). Healthy Leaf exhibits the highest diagonal dominance, with 95.00% of samples correctly classified, and limited confusion primarily with Spot Disease (4.44%). For Rot Disease (stem), 90.95% of samples are correctly identified, with a small proportion misclassified as Healthy Betel Vine (9.05%). Spot Disease maintains a correct classification rate of 91.30%, with minor confusion observed with Healthy Leaf (7.83%).

The class-wise performance metrics presented in Figure 11(b) further quantify robustness under adversarial conditions. Precision values range from 0.90 to 0.95 across all classes, indicating consistent control over false positive predictions. Recall values remain high for all categories, spanning 0.91 to 0.95, demonstrating stable detection of true class instances despite adversarial perturbations. The resulting F1-scores fall within a narrow band between 0.91 and 0.93, reflecting balanced performance between precision and recall for each class.



(a) (b)

Fig. 11. Comparative class-wise performance of the CNN-LSTM model under clean and PGD-adversarial test conditions

Across both representations, no class exhibits disproportionate degradation or performance collapse under adversarial evaluation. The alignment between confusion matrix structure and class-wise metrics confirms that classification performance remains evenly distributed, with errors limited to specific inter-class confusions rather than widespread misclassification. These results provide a consistent empirical profile of the model's behaviour under PGD-based adversarial testing and establish a quantitative foundation for subsequent robustness comparison and discussion.

4.3 Robustness Analysis Summary and Alignment with Research Objectives

Table 2 provides a system-level summary of the proposed CNN-LSTM model's robustness under PGD-based adversarial evaluation. While no prior studies have applied adversarial robustness to betel leaf disease classification, the clean-data performance reported here serves as a baseline. As expected, the accuracy on adversarially perturbed data is lower than that on clean data, confirming the vulnerability of conventional deep learning models to such perturbations and highlighting the need for adversarial training. The model achieves an overall accuracy of 96.96% on clean test data, performance decreases to 92.19% under an iterative PGD attack with perturbation budget $\epsilon = 0.02$, resulting in an absolute degradation of 4.77 percentage points and an accuracy retention of 95.07%. This measurable performance drop empirically confirms RQ1, demonstrating that deep learning image classifiers are vulnerable to visually imperceptible adversarial perturbations when evaluated under strong iterative attacks. At the same time, the limited degradation and high accuracy retention directly address RQ2, indicating that incorporating PGD-based adversarial samples during training significantly improves model stability under adversarial conditions without compromising clean-data performance. Consistent class-wise precision, recall, and F1-score distributions further indicate that robustness gains are not concentrated in specific classes but are evenly maintained across disease categories. Overall, these results demonstrate that the proposed adversarially trained CNN-LSTM model achieves a balanced trade-off between classification accuracy and robustness, supporting its suitability for reliable agricultural disease diagnosis under moderate adversarial perturbations.

Table 2
Robustness Summary of CNN-LSTM under Clean and Adversarial Conditions

Metric	Clean Test Data	PGD Adversarial Test Data	Absolute Accuracy Degradation (pp)
Overall Accuracy (%)	96.96	92.19	4.77
Accuracy Retention (%)	100.00	95.08	–
Attack Type	–	PGD (Iterative)	–
Perturbation Budget (ϵ)	–	0.02 (ℓ∞ norm)	–
Number of PGD Steps	–	80	–
Evaluation Scope	Full test set	Full test set	Consistent

Accuracy retention is defined as the ratio of adversarial accuracy to clean accuracy, expressed as a percentage and it is computed relative to clean test accuracy.

Incorporating PGD-based adversarial training increases the computational requirements of the CNN-LSTM model. Training time is longer due to the additional forward and backward passes for adversarial samples, while inference latency per image remains comparable to the clean-trained model. The saved model size increases from 3.5 MB for the clean-trained model to 10.3 MB for the adversarially trained model, despite the architecture and parameter count remaining unchanged. This increase reflects the storage of adversarial gradients and related training states rather than a change in model complexity, indicating that the approach is feasible for deployment on devices with moderate storage and computational resources.

5. Conclusion and Future Directions

This study examined the vulnerability of deep learning models for betel leaf disease classification to adversarial perturbations and evaluated the effectiveness of adversarial training in improving robustness. Models trained only on clean data suffer measurable performance degradation under visually imperceptible PGD-based attacks, whereas incorporating PGD-generated adversarial samples enables the CNN-LSTM architecture to maintain high clean-data accuracy while limiting degradation under adversarial conditions. Consistent class-wise precision, recall, and F1-scores indicate reliable and balanced predictive behaviour across all disease categories.

While the CNN-LSTM model demonstrates strong performance on the curated betel leaf dataset, its applicability is limited by dataset scale and single-crop focus. The adversarial training strategy is optimized for PGD-style perturbations, which may not capture all real-world variations. Future work should extend validation under diverse field conditions, including mobile-captured images, varying lighting and backgrounds, different disease stages, and sensor noise, and explore larger, multi-crop datasets. Additional robustness strategies and hybrid architectures, alongside explainable AI techniques, may further improve efficiency, interpretability, and practical agricultural applicability.

Acknowledgement
This research was not funded by any grant.

References

[1] Kamilaris, Andreas, and Francesc X. Prenafeta-Boldú. "Deep Learning in Agriculture: A Survey." *Computers and Electronics in Agriculture* 147 (2018): 70–90. https://doi.org/10.1016/j.compag.2018.02.016

[2] Liakos, Konstantinos G., Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. 2018. "Machine Learning in Agriculture: A Review" *Sensors* 18, no. 8: 2674. https://doi.org/10.3390/s18082674

[3] Mohanty, Sharada P., David P. Hughes, and Marcel Salathé. "Using Deep Learning for Image-Based Plant Disease Detection." *Frontiers in Plant Science* 7 (2016): 1419. https://doi.org/10.3389/fpls.2016.01419

[4] Yuan, Xiaoyong, Pan He, Qile Zhu, and Xiaolin Li. 2019. "Adversarial Examples: Attacks and Defenses for Deep Learning." *IEEE Transactions on Neural Networks and Learning Systems* 30 (9): 2805–2824. https://doi.org/10.1109/TNNLS.2018.2886018

[5] Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2574–2582. 2016. https://doi.org/10.1109/CVPR.2016.282

[6] Shi, Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting." In *Proceedings of the 29th International Conference on Neural Information Processing Systems (NeurIPS 2015)*, 802–810. Cambridge, MA: MIT Press, 2015.

[7] Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks." In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[8] Zhu, Xiao Xiang, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources." *IEEE Geoscience and Remote Sensing Magazine* 5, no. 4 (2017): 8–36. https://doi.org/10.1109/MGRS.2017.2762307

[9] Singh, Asheesh, Baskar Ganapathysubramanian, Soumik Sarkar, and Arti Singh. "Deep Learning for Plant Stress Phenotyping: Trends and Future Perspectives." *Trends in Plant Science* 23 (2018). https://doi.org/10.1016/j.tplants.2018.07.004

[10] Barbedo, Jayme. "Factors Influencing the Use of Deep Learning for Plant Disease Recognition." *Biosystems Engineering* 172 (2018). https://doi.org/10.1016/j.biosystemseng.2018.05.013

[11] Shorten, Connor, and Taghi M. Khoshgoftaar. "A Survey on Image Data Augmentation for Deep Learning." *Journal of Big Data* 6, no. 1 (2019): 60. https://doi.org/10.1186/s40537-019-0197-0

[12] Shafahi, Ali, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. "Adversarial Training for Free!" In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019)*. Red Hook, NY: Curran Associates, 2019.

[13] Croce, Francesco, and Matthias Hein. 2020. "Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks." Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR 119:2206-2216.

[14] Ferentinos, Konstantinos P. "Deep Learning Models for Plant Disease Detection and Diagnosis." *Computers and Electronics in Agriculture* 145 (2018): 311–318. https://doi.org/10.1016/j.compag.2018.01.009

[15] Carlini, Nicholas, and David A. Wagner. "Towards Evaluating the Robustness of Neural Networks." In *Proceedings of the IEEE Symposium on Security and Privacy*, 39–57. 2017. https://doi.org/10.1109/SP.2017.49

[16] Esgario, J. G. M., R. A. Krohling, and J. A. Ventura. "Deep Learning for Classification and Severity Estimation of Coffee Leaf Biotic Stress." *Computers and Electronics in Agriculture* 169 (2020): 105162. https://doi.org/10.1016/j.compag.2019.105162

[17] Balram, G., and K. Kiran Kumar. "Crop Field Monitoring and Disease Detection of Plants in Smart Agriculture Using Internet of Things." *International Journal of Advanced Computer Science and Applications* 13, no. 7 (2022). https://doi.org/10.14569/IJACSA.2022.0130795

[18] Nazir, Ahsan, Jingsha He, Nafei Zhu, Saima Qureshi, Siraj Qureshi, Ahsan Wajahat, and Muhammad Salman Pathan. "A Deep Learning-Based Novel Hybrid CNN-LSTM Architecture for Efficient Detection of Threats in the IoT Ecosystem." *Ain Shams Engineering Journal* 15 (2024): 102777. https://doi.org/10.1016/j.asej.2024.102777

[19] van der Velden, Bas H.M., Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. "Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis." *Medical Image Analysis* 79 (2022): 102470. https://doi.org/10.1016/j.media.2022.102470

[20] Gardezi, Maaz, Bhavna Joshi, Donna Rizzo, Mark Ryan, Edward Prutzer, Skye Brugler, and Ali Dadkhah. "Artificial Intelligence in Farming: Challenges and Opportunities for Building Trust." *Agronomy Journal* 116 (2023). https://doi.org/10.1002/agj2.21353

[21] Silwal, Abhisesh, Tanvir Parhar, Francisco Yandun, Harjatin Baweja, and George Kantor. "A Robust Illumination-Invariant Camera System for Agricultural Applications." In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3292–3298. 2021. https://doi.org/10.1109/IROS51168.2021.9636542

[22] Mahmoud, Nesma, Indrek Virro, A. Zaman, Tormi Lillerand, Wai Chan, Olga Liivapuu, Kallol Roy, and Jüri Olt. "Robust Object Detection Under Smooth Perturbations in Precision Agriculture." *AgriEngineering* 6, no. 4 (2024): 4570–4584. https://doi.org/10.3390/agriengineering6040261

[23] Rajarajeswari, Perepi, Vijaya Redrowthu, Yelamanchi Jahnavi, Vasavi Ravuri, Sampath Alankritha, and Maddukuri Vani, Rajesh Muthu, Silvia Gaftandzhieva. "Performance Evaluation of Generative Adversarial Networks for Anime Face Synthesis Using Deep Learning Approaches." *Multidisciplinary Science Journal* 7 (2024): 2025093. https://doi.org/10.31893/multiscience.2025093

[24] Guan, Shuyue, and Murray Loew. "Evaluation of Generative Adversarial Network Performance Based on Direct Analysis of Generated Images." In *2019 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–5. 2019. https://doi.org/10.1109/AIPR47015.2019.9174595

[25] Park, Hyejin, Taaha Waseem, Wen Qi Teo, Ying Hwei Low, Mei Kuan Lim, and Chun Yong Chong. "Robustness Evaluation of Stacked Generative Adversarial Networks Using Metamorphic Testing." In *Proceedings of the 2021 IEEE/ACM 6th International Workshop on Metamorphic Testing (MET)*, 1–8. 2021. https://doi.org/10.1109/MET52542.2021.00008

[26] Alexander, Carrie S., Mark Yarborough, and Aaron Smith. "Who Is Responsible for 'Responsible AI'?: Navigating Challenges to Build Trust in AI Agriculture and Food System Technology." *Precision Agriculture* 25, no. 1 (2024): 146–185. https://doi.org/10.1007/s11119-023-10063-3

[27] Dhilipkumar, V., S. DineshKumar, S. Maheswari, and Esra Sipahi Döngül. "Trustworthy AI Knowledge Systems for Precision Agriculture and Agribusiness Management: Detecting Crop Health Using Satellite Imagery." *Science Talks* 16 (2025): 100494. https://doi.org/10.1016/j.sctalk.2025.100494

[28] David, Femi, and Manapakkam Anandan Mukunthan. "Betel Leaf Diseases Classification using Machine Learning Algorithm: A Feasible Approach." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 40, no. 1 (2024): 74–86. https://doi.org/10.1234/jaraset.2024.40.1.74

[29] Muhammad Hanif Tunio, Jian Ping Li, Xiaoyang Zeng, Awais Ahmed, Syed Attique Shah, Hisam-Uddin Shaikh, Ghulam Ali Mallah, and Imam Abdullahi Yahya. "Advancing Plant Disease Classification: A Robust and Generalized Approach with Transformer-Fused Convolution and Wasserstein Domain Adaptation." *Computers and Electronics in Agriculture* 227 (2024): 109574. https://doi.org/10.1016/j.compag.2024.109574

[30] Dablain, Damien, Kristen N. Jacobson, Colin Bellinger, Mark Roberts, and Nitesh V. Chawla. "Understanding CNN Fragility When Learning with Imbalanced Data." *Machine Learning* 113, no. 7 (2024): 4785–4810. https://doi.org/10.1007/s10994-023-06326-9

[31] Zago, João G., Eric A. Antonelo, Fabio L. Baldissera, and Rodrigo T. Saad. "Benford's Law: What Does It Say on Adversarial Images?" *Journal of Visual Communication and Image Representation* 93 (2023): 103818. https://doi.org/10.1016/j.jvcir.2023.103818

[32] Ma, Linhai, and Liang Liang. "Towards Lifting the Trade-Off Between Accuracy and Adversarial Robustness of Deep Neural Networks with Application on COVID-19 CT Image Classification and Medical Image Segmentation." In *Proceedings*, 67. 2023. https://doi.org/10.1117/12.2653392

[33] Khagram, Tanisha. "Performance Benchmarking of CNN Architectures for Fine-Grained Image Classification on CIFAR-100." *International Journal for Research in Applied Science and Engineering Technology* 13 (2025): 2657–2662. https://doi.org/10.22214/ijraset.2025.75713

[34] Zhang, Kai, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising." *IEEE Transactions on Image Processing* 26, no. 7 (2017): 3142–3155. https://doi.org/10.1109/TIP.2017.2662206

[35] Latif, Ghazanfar, Jaafar Alghazo, Majid Ali Khan, Ghassen Ben Brahim, Khaled Fawagreh, and Nazeeruddin Mohammad. "Deep Convolutional Neural Network (CNN) Model Optimization Techniques—Review for Medical Imaging." *AIMS Mathematics* 9, no. 8 (2024): 20539–20571. https://doi.org/10.3934/math.2024998

[36] Zhou, Kang, Huyin Zhang, and Fei Li. "TransNav: Spatial Sequential Transformer Network for Visual Navigation." *Journal of Computational Design and Engineering* 9, no. 5 (2022): 1866–1878. https://doi.org/10.1093/jcde/qwac084

[37] Feng, Shiyang, Tianyue Chen, and Hao Sun. "Long Short-Term Memory Spatial Transformer Network." In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, 239–242. 2019. https://doi.org/10.1109/ITAIC.2019.8785574

[38] Sun, Hao, Yanjie Xu, Gangyao Kuang, and Jin Chen. "Adversarial Robustness Evaluation of Deep Convolutional Neural Network Based SAR ATR Algorithm." In *2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 5263–5266. 2021. https://doi.org/10.1109/IGARSS47720.2021.9554783