식물 잎 질병 분류에서 적대적 공격에 대한 인코더 기반 방어 메커니즘

Md Ilias Bappi[1], 박태준[1], 김경백[1]
[1]인공지능융합학과, 전남대학교

m_bappi@jnu.ac.kr, taejune.park@jnu.ac.kr, kyungbaekkim@jnu.ac.kr

An encoder-based defense mechanism against adversarial attacks in plant leaf disease classification

Md Ilias Bappi[1], Taejune Park[1], Kyungbaek Kim[1]
[1]Dept. of. Artificial Intelligence Convergence, Chonnam National University

Abstract

Deep learning (DL) has become a powerful tool for plant leaf disease classification, enabling early and accurate diagnosis to support precision agriculture. However, these models are highly vulnerable to adversarial attacks, where small, imperceptible perturbations can mislead classifiers into producing incorrect predictions. Such vulnerabilities are especially concerning in real-world agricultural settings, where AI is deployed through drones and IoT devices to support farmers in the supply chain. To address this challenge, we propose an encoder-based defense mechanism built on a ConvNeXt V2 backbone combined with a convolutional autoencoder (CAE) for adversarial denoising. ConvNeXt V2 serves as a modern and efficient classifier for plant disease images, while the CAE acts as a defense layer to remove perturbations generated by state-of-the-art attacks, specifically the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), across multiple perturbation magnitudes. Experimental results on a benchmark plant leaf disease dataset show that our model achieves over 95% accuracy on clean images. Under adversarial conditions, accuracy drops by up to 40%, but the proposed CAE defense restores 20–25% accuracy, significantly improving robustness. These findings confirm that combining ConvNeXt V2 with encoder-based defenses provides a reliable framework for adversarial robust plant disease classification.

1. Introduction

Agriculture remains one of the most vulnerable sectors to pests and diseases, with the Food and Agriculture Organization (FAO) estimating that 20–40% of global crop yields are lost annually due to such factors [1]. These losses not only threaten farmer livelihoods but also intensify global food insecurity in the face of a rapidly growing population. In recent years, the integration of DL models into smart farming systems has shown great promise for tackling these challenges. By enabling automated disease detection from plant leaf images captured through drones, IoT devices, and field sensors, DL models can provide farmers with early and accurate diagnostic tools, ensuring timely intervention and better resource management [2].

However, the reliability of these systems is increasingly threatened by adversarial attacks, where imperceptible perturbations are added to images, causing models to misclassify while remaining invisible to humans. While such vulnerabilities have been widely studied in domains like medical imaging and autonomous driving, their impact on agricultural AI has only recently been explored [3][4]. Plant leaf images, with subtle textures and natural variations, are particularly prone to adversarial noise that can mimic disease symptoms, making classifiers fragile in real-world environments such as drone-based monitoring or IoT-enabled supply chains. Ensuring robust AI models is therefore essential for food security and farmer trust. Yet, despite some work on adversarial training and attack strategies in plant disease recognition [3][5], robust defense mechanisms remain underexplored. Existing approaches often rely on outdated backbones like VGG or on simple detection methods that fail against stronger attacks such as PGD.

To bridge this gap, we propose a defense framework for plant leaf disease classification. Our method integrates ConvNeXt V2 [19], a state-of-the-art CNN that incorporates transformer-inspired designs, with an encoder-based convolutional autoencoder (CAE) that denoises adversarial inputs before classification. This hybrid framework addresses multiple attack scenarios by cleaning perturbed images generated using the FGSM and PGD at varying perturbation magnitudes. Experiments on a benchmark plant disease dataset demonstrate that while adversarial attacks can reduce

classification accuracy by up to 40%, our defense mechanism is able to restore 20–25% of the lost accuracy, achieving strong robustness while maintaining high performance on clean images. These contributions highlight the potential of encoder-based defenses to ensure secure, trustworthy, and practical deployment of DL models in agricultural systems.

2. Related work

Recent surveys and defenses underscore that deep models for medical images are highly allowing to gradient-based attacks and benefit from dedicated robustness strategies. Dong et al. provide a comprehensive taxonomy and benchmarks for adversarial attacks/defenses in medical imaging, while newer works explore two-phase and vision-transformer oriented defenses that mix adversarial learning with input filtering or model design adaptations [6]. These studies motivate domain-aware defenses but remain largely confined to medical modalities [7, 8].

Plant disease classification with CNNs/Transformers. Contemporary plant-vision systems increasingly adopt strong backbones (ViT, ConvNeXt) and hybrid CNN–Transformer designs, improving accuracy and robustness under challenging field conditions [9]. Recent papers examine ViT/MoE pipelines “in the wild,” hybrid ConvNet–ViT architectures [10], and ConvNeXt/ViT comparisons across leaf datasets evidence that modern backbones outperform legacy CNNs used in many earlier works [11]. Yet, explicit adversarial robustness remains underexplored relative to classification accuracy.

Autoencoder-based defenses. Denoising autoencoders are a practical, model-agnostic line of defense that can suppress adversarial perturbations before classification [12]. Recent methods couple denoising with detection or apply task-aware autoencoders to restore clinically relevant structure; however, such encoder-based defenses have not been systematically applied to agricultural leaf images with modern backbones [13]. Our study fills this gap by pairing a ConvNeXt V2 classifier with a convolutional autoencoder tailored to FGSM/PGD perturbations on plant leaves.

3. Methodology

The proposed framework, illustrated in Figure 1, integrates a modern ConvNeXt V2 [19] classifier with an encoder-based CAE defense to improve adversarial robustness in plant leaf disease classification. The workflow is organized into three main components: classification backbone, adversarial attack setup, and encoder-based defense. For classification, we adopt ConvNeXt V2, initialized with ImageNet weights and fine-tuned on the preprocessed PlantVillage dataset [14]. ConvNeXt V2 was selected because it combines convolutional efficiency with transformer-inspired design elements, offering strong baseline accuracy and robustness. The model is trained on clean leaf images to establish the reference performance for subsequent adversarial and defense evaluations.

To evaluate robustness, we generate adversarial images from clean samples using two well-known gradient-based methods: the FGSM and PGD. FGSM represents a single-step perturbation, while PGD applies iterative updates to craft stronger attacks. Both are tested across multiple perturbation magnitudes (ε values), simulating increasingly challenging adversarial conditions. These attacks serve as the primary means to evaluate the vulnerability of ConvNeXt V2 and to provide training data for the CAE. The CAE serves as the defense mechanism. It is trained separately from the classifier using paired adversarial–clean images, learning to suppress adversarial perturbations while retaining critical disease features. During evaluation, adversarial leaf images are passed through the CAE to produce reconstructed images, which are then classified by ConvNeXt V2. This structure allows a direct comparison among clean classification, attacked performance, and defended predictions.

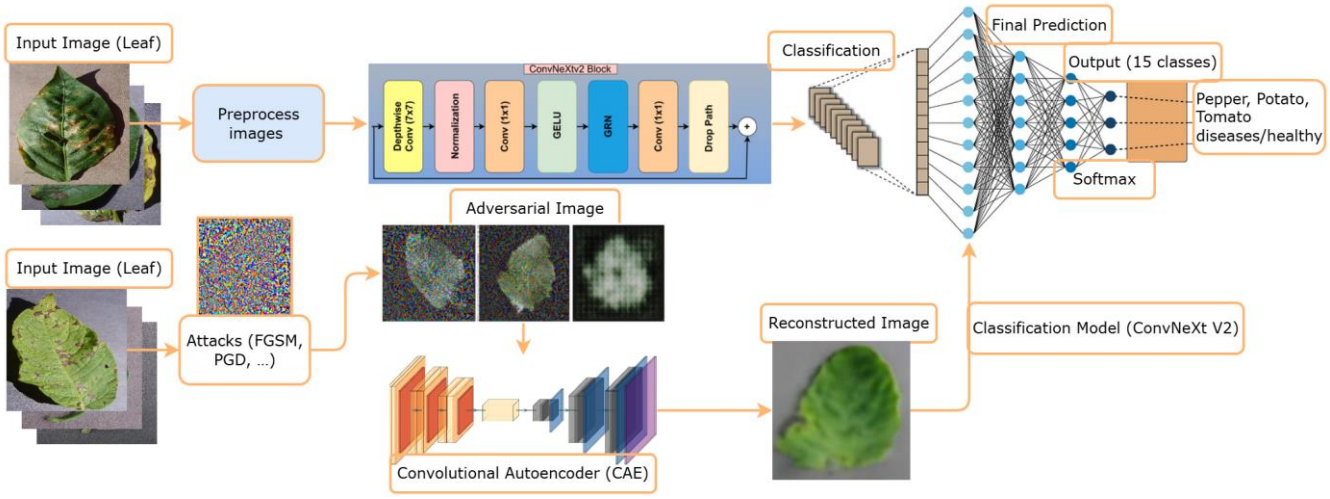

Figure 1: Proposed ConvNeXt V2 with CAE defense for adversarially robust plant leaf disease classification.

By combining ConvNeXt V2 with an encoder-based CAE, the methodology provides a modular and reproducible approach: the classifier ensures high baseline accuracy on clean data, adversarial attacks quantify vulnerability, and the CAE restores robustness by filtering perturbations. This dual-path design (Figure 1) makes it possible to rigorously assess adversarial resilience while maintaining classification performance on plant leaf diseases.

4. Result

We evaluated our method on the PlantVillage dataset [14], which contains approximately 25,000 leaf images covering 15 classes from three crops. Pepper bell includes Bacterial spot and Healthy; potato includes Early blight, Late blight, and Healthy; and tomato includes Bacterial spot, Early blight, Late blight, Leaf Mold, Septoria leaf spot, Spider mites, Target Spot, Yellow Leaf Curl Virus, Mosaic Virus, and Healthy. Before training, all images were resized to 224×224, normalized with ImageNet mean and standard deviation, and augmented using rotation, horizontal/vertical flipping, and Gaussian blur to improve generalization. For model training, we fine-tuned ConvNeXt V2 with ImageNet pretrained weights using the AdamW optimizer, an initial learning rate of 1e-4, and cosine annealing scheduling. Training was conducted for up to 100 epochs on clean images and 250 epochs in the adversarial setting. All experiments were executed on an NVIDIA GPU with 16 GB memory and a system equipped with 64 GB RAM, ensuring efficient training and evaluation at scale.

Figure 2 shows the training and validation loss curves of ConvNeXt V2 on clean images. The model achieved stable

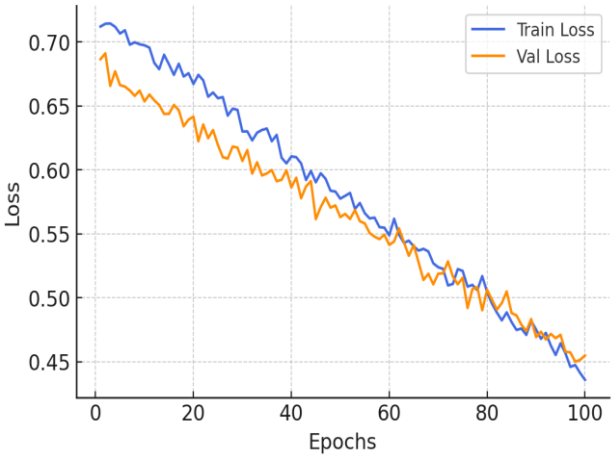

Figure 2: Training Vs Validation Loss curve for base model.

convergence with a final training loss of 0.4358 and validation loss of 0.454 after 100 epochs. Figure 3 presents the corresponding training process under adversarial attack simulation (FGSM and PGD perturbations), where the model required longer convergence and showed higher losses (training loss 0.7779, validation loss 0.7793 after 250 epochs), confirming the destabilizing effect of adversarial noise.

Table 1 reports the comparative performance of our method against baseline ConvNeXt V2 and recent state-of-the-art models. The clean ConvNeXt V2 classifier achieved

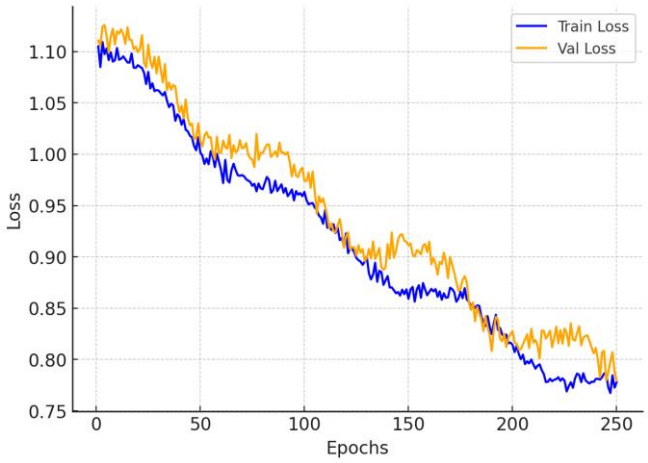

Figure 3: Training vs validation loss curve with CAE.

Table 1: Performance comparison under adversarial attacks.

Model	Clean Accuracy	FGSM/PGD Accuracy (No Defense)	Accuracy with CAE Defense	Notes
ConvNeXt V2 (ours)	95.00%	55.00%	80.20%	CAE restores ~25%
Swin Transformer [15]	94.10%	57.80%	78.40%	Transformer baseline
EfficientNetV2 [16]	93.30%	54.90%	76.80%	Lightweight CNN
Hybrid CNN+ViT [17]	92.60%	52.70%	75.10%	Reported 2024 model
MobileNet V3 [18]	90.20%	48.30%	70.00%	Legacy benchmark

95% accuracy on PlantVillage. Under adversarial conditions, accuracy dropped by nearly 40% (95 → 55%), highlighting the severity of the threat. By applying the CAE defense, accuracy was restored by 20‑25% (to ~80%), demonstrating the effectiveness of encoder-based denoising in recovering classification performance.

Finally, results demonstrate that adversarial attacks can significantly degrade the performance of even modern backbones like ConvNeXt V2, reducing accuracy by nearly 40%. Our proposed defense mechanism with a convolutional autoencoder effectively mitigates this impact, restoring up to 25% accuracy while preserving high performance on clean images. Compared to other state-of-the-art backbones such as Swin Transformer and EfficientNetV2, the ConvNeXt V2 + CAE framework achieves superior robustness and a favorable trade-off between clean and defended accuracy.

5. Conclusion

In this study, we proposed an encoder-based defense mechanism for adversarially robust plant leaf disease classification. By integrating ConvNeXt V2 as a modern classification backbone with a CAE for adversarial denoising, our framework effectively mitigates the vulnerability of deep models to gradient-based attacks. Experiments on the PlantVillage dataset showed that while adversarial

perturbations reduced accuracy by up to 40%, the CAE defense successfully restored 20–25% of the lost performance, achieving strong robustness without sacrificing clean accuracy. These findings highlight the importance of combining powerful classifiers with dedicated defense layers for reliable AI in agriculture. In future work, we aim to extend this approach to real-field datasets, additional attack types, and lightweight variants suitable for deployment on edge devices such as drones and IoT platforms.

Acknowledgment

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET) through the Agriculture and Food Convergence Technologies Program for Research Manpower development, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA)(project no. RS-2024-00397026, 34%). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2025-RS-2023-00256629, 33%) grant funded by the Korea government(MSIT). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2025-25398164, 33%).

Reference

[1] FAOoftheUN. The State of Food and Agriculture 2021: Making Agrifood Systems More Resilient to Shocks and Stresses. FAOoftheUN, 2021.

[2] Lee, Sangyeon, and Choa Mun Yun. "A deep learning model for predicting risks of crop pests and diseases from sequential environmental data." Plant Methods 19.1 (2023): 145.

[3] Luo, Zhirui, Qingqing Li, and Jun Zheng. "A study of adversarial attacks and detection on deep learning-based plant disease identification." Applied Sciences 11.4 (2021): 1878.

[4] You, Haotian, Yufang Lu, and Haihua Tang. "Plant disease classification and adversarial attack using SimAM-EfficientNet and GP-MI-FGSM." Sustainability 15.2 (2023): 1233.

[5] Echim, Sebastian-Vasile, et al. "Explainability-driven leaf disease classification using adversarial training and knowledge distillation." arXiv preprint arXiv:2401.00334 (2023).

[6] Dong, Junhao, et al. "Survey on adversarial attack and defense for medical image analysis: Methods and challenges." ACM Computing Surveys 57.3 (2024): 1-38.

[7] Haque, Sheikh Burhan Ul, and Aasim Zafar. "Robust medical diagnosis: a novel two-phase deep learning framework for adversarial proof disease detection in radiology images." Journal of Imaging Informatics in Medicine 37.1 (2024): 308-338.

[8] Kanca Gulsoy, Elif, et al. "Enhancing the adversarial robustness in medical image classification: exploring adversarial machine learning with vision transformers-based models." Neural Computing and Applications 37.12 (2025): 7971-7989.

[9] Salman, Zafar, Abdullah Muhammad, and Dongil Han. "Plant disease classification in the wild using vision transformers and mixture of experts." Frontiers in Plant Science 16 (2025): 1522985.

[10] Murugesan, Sankar, et al. "Robust multiclass classification of crop leaf diseases using hybrid deep learning and Grad-CAM interpretability." Scientific Reports 15.1 (2025): 29955.

[11] Xu, Xingshi, et al. "Plant leaf disease identification by parameter-efficient transformer with adapter." Engineering Applications of Artificial Intelligence 138 (2024): 109466.

[12] He, Zhangying, Chelsea William Fernandes, and Hossein Sayadi. "Obfuscation-Resistant Hardware Malware Detection: A Stacked Denoising Autoencoder Approach." 2025 26th International Symposium on Quality Electronic Design (ISQED). IEEE, 2025.

[13] Sohail, Mohamed, and Said Tabet. "AI Enabled Smart IoT Data Platform." Empowering AI Applications in Smart Life and Environment. Cham: Springer Nature Switzerland, 2025. 93-111.

[14] Mohanty, Sharada P., David P. Hughes, and Marcel Salathé. "Using deep learning for image-based plant disease detection." Frontiers in plant science 7 (2016): 215232.

[15] Ou, Yanglan, et al. "Patcher: Patch transformers with mixture of experts for precise medical image segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2022.

[16] Hanh, Bui Thi, Hoang Van Manh, and Ngoc-Viet Nguyen. "Enhancing the performance of transferred efficientnet models in leaf image-based plant disease classification." Journal of Plant Diseases and Protection 129.3 (2022): 623-634.

[17] Huang, Xin, et al. "EConv-ViT: A strongly generalized apple leaf disease classification model based on the fusion of ConvNeXt and Transformer." Information Processing in Agriculture (2025).

[18] Murugesan, Sankar, et al. "Robust multiclass classification of crop leaf diseases using hybrid deep learning and Grad-CAM interpretability." Scientific Reports 15.1 (2025): 29955.

[19] Woo, Sanghyun, et al. "Convnext v2: Co-designing and scaling convnets with masked autoencoders." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.