Adversarial Image Detection Using Deep Learning in Agricultural Contexts

Md Nazmul Kabir Sikder[1], Mehmet Oguz Yardimci[2], Trey Ward[3], Shubham Laxmikant Deshmukh[4], and Feras A. Batarseh[5]

[1]Virginia Tech, Commonwealth Cyber Initiative (CCI), Arlington, VA, 22203, USA
[2,4]Virginia Tech, Department of Computer Science, Arlington, VA, 22203, USA
[3,5]Virginia Tech, Department of Biological Systems Engineering, Arlington, VA, 22203, USA

ABSTRACT

The gradual digitalization of agricultural systems through data-driven techniques has reshaped production growth. However, this transformation has also introduced new vulnerabilities, exposing these systems to cyber threats. While numerous domain-specific attack detection methods have been proposed, there is a lack of comprehensive cybersecurity frameworks tailored for agriculture, particularly as AI becomes increasingly integrated into these systems. To address this gap, we propose a novel framework capable of classifying high-fidelity adversarial plant images. This supervised approach not only detects attacks but also able to identify their specific source models. We employ state-of-the-art GAN architectures, including StyleGAN2 and StyleGAN3, alongside powerful diffusion models such as DS8, BLIP, and Pix2Pix, to produce diverse adversarial images via both image-to-image and text-to-image generation. These images are then used to train a classifier capable of distinguishing among all generation classes. Our experiments include comparative classification tasks, and logarithmic accuracy degradation with increasing class count. This demonstrates the scalability of the framework, allowing additional computer vision tasks to be incorporated without compromising performance. As GAN and diffusion models continue to advance, our framework is designed to evolve, ensuring its generation and detection capabilities remain robust against emerging threats.

1 Introduction and Motivation

By 2050, global food production must increase by approximately 70% to sustain a projected world population nearing 10 billion[1,2]. To meet this challenge, agriculture is undergoing rapid digital transformation, widely adopting smart technologies, Internet of Things (IoT), artificial intelligence (AI), and cloud-based infrastructure, collectively termed *Agriculture 4.0*[3]. The expansion of smart agriculture is significantly driven by increased integration of connected devices and intelligent analytics platforms, optimizing yields, conserving resources, and enhancing overall productivity[4].

However, this rapid digitization also expands the threat surface of agricultural cyber-physical systems (CPS), making them increasingly attractive targets for cyber adversaries. While traditional risks such as climate instability, extreme weather, and pest or disease outbreaks persist[5–7], new vulnerabilities have emerged from interconnected IoT networks, cloud platforms, and AI-driven decision-making systems[8,9]. Cyber-physical threats, ranging from data manipulation to adversarial image injection, could cause supply chain disruptions, economic loss, compromised food safety, and risks to public health[10]. For instance, cyberattacks in this domain may manifest as: (1) a denial-of-service (DoS) attack on temperature or humidity sensors that disrupts smart irrigation schedules and leads to crop stress or failure[11]; (2) spoofing of soil pH and conductivity sensor data, tricking AI systems into recommending harmful fertilizer or pesticide usage[12]; and (3) GPS spoofing of Unmanned Aerial Vehicles (UAVs) or tampering with fluid level sensors in hydroponic systems, causing nutrient imbalance or physical crop damage[13].

A particularly concerning vector is the malicious use of advanced generative models, such as GANs, image diffusion models, and adversarial perturbation techniques, to deceive AI vision systems. For example, the *RisingAttacK* framework demonstrates how imperceptible pixel-level perturbations can mislead widely used models like ResNet-50, DenseNet-121, and Vision Transformers (ViT)[14]. This capability to manipulate what an AI sees poses significant risks for agriculture, where visual data inform yield estimation, disease detection, and quality control. In smart agriculture, UAVs equipped with RGB, multispectral, and hyperspectral cameras, combined with DL-based models, are routinely deployed for crop disease detection, phenotyping, weed identification, and pest monitoring[15–17]. Adversarial manipulation of these image streams could therefore result in missed detection of pathogens, false yield predictions, or improper pesticide spraying, ultimately leading to large-scale economic loss and food insecurity.

Previous work by Alferidah and Algosaibi[18] explored this possibility specifically for agricultural systems through a diffusion-based noise attack directed at images, videos, audios, and text data being processed by a classification model. The

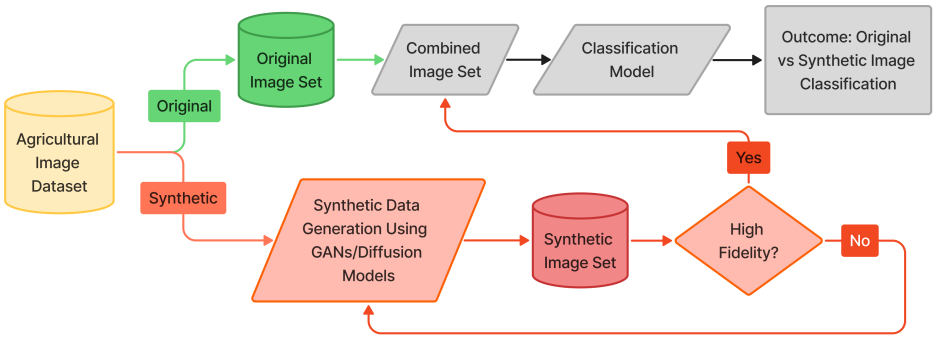


Figure 1. Pipeline for detecting original vs. synthetic agricultural images. Original and synthetic (GAN/ diffusion/ Pix2Pix) images feed a classifier that predicts *Original* or *Synthetic*.

misclassifications caused by the example attack highlight the importance and capabilities of systems similar to what was developed by Salman et al.[19]. The immunization process described prevents effective AI-driven manipulation of images by inducing unrealistic alterations, commonly referred to as hallucinations. However a lingering question remains regarding attribution, necessitating the development of classification models capable of identifying various generative AI techniques that malicious actors may employ to carry out a cyber-attack.

Additionally, the experiment gives credence to GAN-based image augmentation capabilities in data-poisoning attacks targeting the underlying training data supporting classification models needed to detect potential cyber-attacks. As shown by Rahmen Et Al. [20] GANs can play a role in generating augmented data needed to train classification models in farming contexts due to the lack of available data. Ghazal Et Al. [21] documents the growing application and use of computer vision systems in farming, highlighting the importance of safeguarding training data sets with a classification model capable of detecting how an image was altered (i.e. Diffusion or GANS) to signal a potential attack

Motivated by this threat landscape, we present a novel adversarial pipeline (Figure 1) designed specifically for agricultural CPS. Our solution leverages state-of-the-art GANs and diffusion models, including StyleGAN2[22], StyleGAN3[23], Pix2Pix[24], Dreamshaper-8[25], and BLIP[26] and synthesizes high-fidelity, domain-specific adversarial plant images. Plant image classes selected (Tomato, Apple, & Maize) represent three distinct cultivation methods and crop types with substantial risk of an attack targeting computer vision systems. The tomato class represents fruiting plants commonly found in greenhouses and other forms of controlled environment agriculture (CEA); Apples represent fruits commonly cultivated outdoors in orchards or vineyards, & Maize represents a field crop cultivated on industrial scales as a key part of rural economies. Additionally these crops each have 1000 healthy & unhealthy publicly available images, in common formats within classes, enabling experimentation.

These images are then used to train a robust classifier capable of both binary detection and fine-grained source (GANs vs Diffusion models) attribution. We demonstrate through extensive cross-model and cross-generation experiments that our framework maintains high classification accuracy, even as the number of generative model classes increases, thus highlighting its scalability and adaptability for evolving threats.

1.1 Related Works and Contribution

Cybersecurity is a growing concern as food production systems incorporate image data and computer vision technologies into their processes[27,28]. Despite significant advances in image forgery and deepfake detection, most existing research targets general-purpose domains such as human faces, artwork, or multimedia content. Comparable efforts in agriculture are rare, even though generative models can produce synthetic plant imagery realistic enough to bypass AI-based decision systems. This lack of domain-specific frameworks leaves agricultural CPS vulnerable to sophisticated content-based attacks.

Generative Adversarial Networks (GANs)[29] and diffusion models[30] have revolutionized synthetic image generation. Architectures such as StyleGAN2[22] and StyleGAN3[23] achieve photorealism with fine-grained control, while methods such as Stable Diffusion[25], Pix2Pix diffusion[24], and BLIP-based pipelines[26] excel at semantic alignment and texture fidelity. Their increasing realism, however, presents new security challenges across industries.

Detection techniques have evolved from handcrafted features to deep learning-based approaches. Convolutional Neural Networks(CNN)-based detectors[31,32] exploit spatial artifacts, hybrid CNN–RNN models[33] incorporate temporal dependencies, and frequency-aware methods[34] improve cross-model generalization. Multi-branch architectures integrating RGB and spectral cues[35] enhance robustness, while multimodal methods such as AntifakePrompt[36] and DE-FAKE[37] combine semantic and visual signals for detection.

Nevertheless, challenges remain: overfitting to specific generators limits generalization[38]; most datasets are human-centric; and multi-class source attribution is underexplored despite its operational importance[31]. In agriculture, where visual AI models

directly influence decision-making, these gaps present a pressing need for tailored solutions.

Research at the intersection of agriculture and AI security has largely examined *adversarial perturbations* against plant-image classifiers and defenses via adversarial training. Luo et al.[39] systematized white-box attacks and simple detectors on plant disease identification models, while You et al.[40] and Li and Lu[41] proposed stronger *MI-FGSM* variants to evaluate robustness. Echim et al.[42] and Yang et al.[43] explored adversarial training strategies, demonstrating improved resilience in disease classification pipelines.

Parallel efforts have leveraged *generative models*, particularly GANs, primarily for data augmentation rather than attack simulation. Nazki et al.[44], Bi and Hu[45], and Wang et al.[46] applied GAN-based image translation, augmentation, or hyperspectral data generation to strengthen plant disease detection. More recent works include CycleGAN-based pear disease recognition[47] and GAN-augmented vision transformers for leaf classification[48]. While effective for boosting accuracy, these studies do not explicitly address content-based attacks.

Recent work has shown that diffusion models can be exploited to craft adversarial examples that are both imperceptible to humans and transferable across model architectures[49]. These attacks subtly manipulate the generative process itself, rather than adding noise post hoc, resulting in semantically coherent yet deceptive images. In agricultural contexts, such methods can hypothetically be used to generate synthetic plant imagery that misleads disease classifiers or health status detectors, posing a novel threat to CPS reliant on visual data.

By contrast, the *forensics literature* in computer vision highlights that GANs and diffusion models leave learnable "fingerprints," enabling both detection and *source attribution* across multiple generators[38,50,51]. However, such approaches focus on human faces or open-domain images rather than agricultural CPS.

Our work differs by introducing a unified, agriculture-tailored framework that (i) generates high-fidelity, adversarial-style plant images using both GAN and diffusion models (image-to-image and text-to-image), and (ii) detects and attributes them across multiple generators, scaling from binary to multi-class without degradation in performance. To our knowledge, prior agricultural studies have used GANs chiefly for data augmentation or pixel-level perturbation analysis, but none integrate *cross-model content forgery generation with multi-class attribution* for agricultural CPS cybersecurity. This capability not only advances detection and attribution research within agricultural contexts and lays the groundwork for defending these systems against future cyber-attacks involving synthetic imagery.

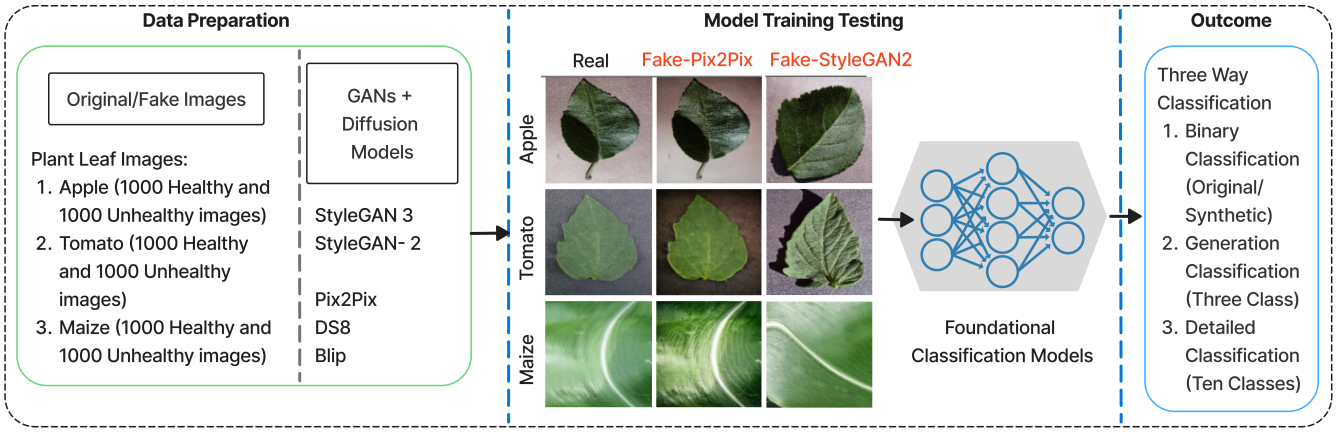


Figure 2. Overview of the synthetic leaf image generation and classification framework.

This paper makes the following key contributions:

1. Designing the first unified adversarial image generation and detection framework specifically for agricultural CPS, integrating multiple state-of-the-art GAN and diffusion models.

2. Generating a large-scale, high-fidelity synthetic dataset of plant images (including healthy and unhealthy categories) for adversarial detection research.

3. Developing a scalable classifier capable of both binary real/synthetic detection and fine-grained source attribution across multiple generative models.

1.2 Research Questions

This study is guided by the following research questions, which emerge from the dual challenges of high-fidelity synthetic image generation and robust detection in agricultural cyber-physical systems:

1. **RQ1 Classification Granularity:** How does performance change as classification moves from binary (healthy vs. unhealthy) to generation-source attribution (original vs. GAN vs. diffusion) and finally to detailed multi-class attribution (10-way crop–health–generator categories)? *Motivation:* Understanding scalability across tasks of increasing granularity is essential to determine whether detection models can maintain reliability when confronted with more complex attribution requirements.

2. **RQ2 Crop-Specific Challenges:** Are certain plant types (e.g., maize vs. tomato vs. apple) inherently more difficult for synthetic image detection due to texture complexity or subtle disease patterns, and do CNNs and transformers respond differently to these challenges? *Motivation:* Agricultural datasets vary significantly in visual texture and disease manifestation, making it important to assess whether model architectures generalize consistently across crops or reveal vulnerabilities tied to biological variation.

3. **RQ3Adversarial Scenarios:** How could synthetic images be exploited to conceal real outbreaks (healthy-look for diseased leaves) or fabricate false epidemics (diseased-look for healthy leaves), and can detection models serve as effective safeguards against such attacks? *Motivation:* Beyond classification accuracy, evaluating the potential misuse of generative models highlights the real-world stakes of agricultural cyber-biosecurity, where manipulated images could directly impact disease monitoring and food security.

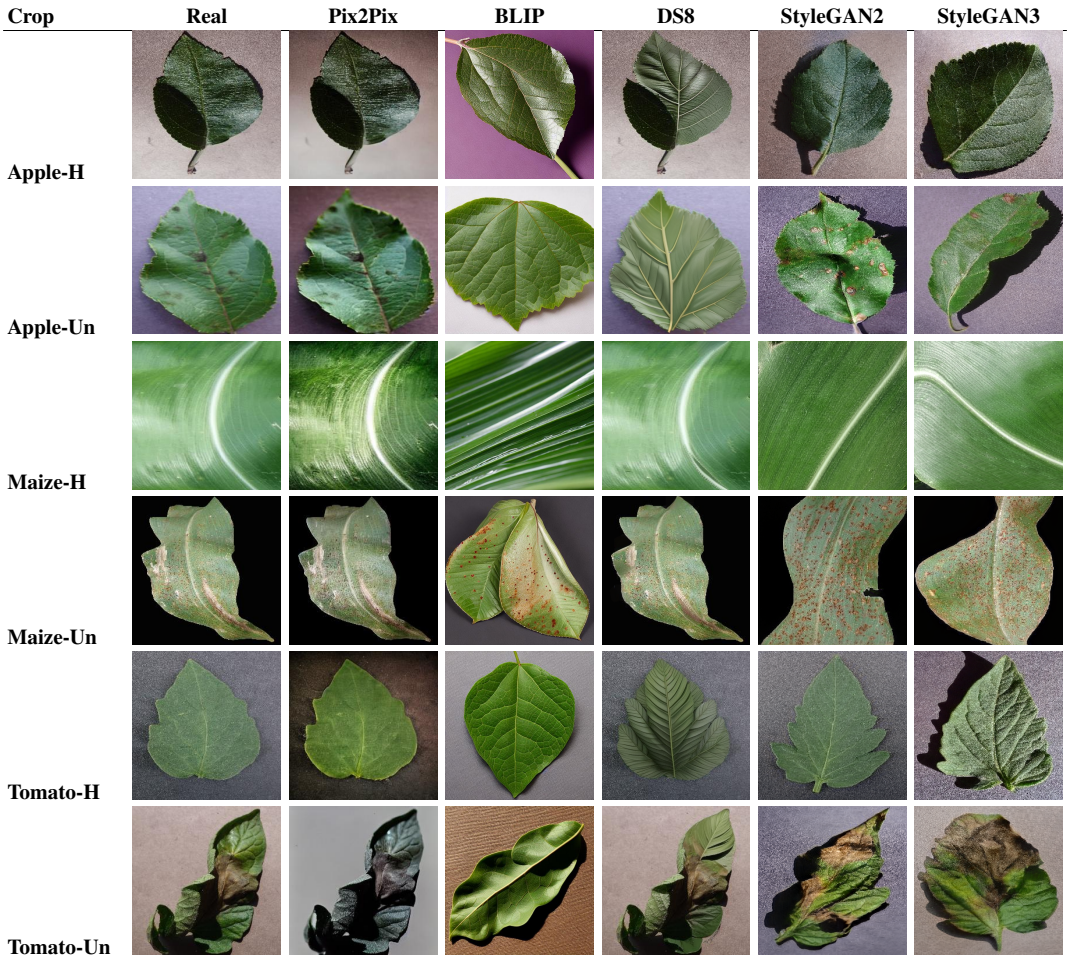


Table 1. Comparison of original and synthetic images (Pix2Pix, BLIP, DS8, StyleGAN2, StyleGAN3) across crop types and health conditions. H = healthy leaf, Un = unhealthy leaf.

2 Methods

In this study, we evaluate both original and synthetically generated leaf images to support downstream analysis and model training. Table 1 presents a comparative grid of representative samples across three crop types (apple, maize, and tomato) and

two health conditions (healthy and unhealthy). Each row shows the real leaf alongside synthetic variants produced by Pix2Pix, BLIP, DS8, StyleGAN2, and StyleGAN3. This visualization highlights differences in texture, color fidelity, and structural consistency across generation techniques. In the following subsections, we describe in detail the process used to generate these synthetic images.

2.1 Synthetic Data Generation via GANs

Our synthetic data pipeline evolved through multiple stages, beginning with exploratory trials of conditional GANs and culminating in high-fidelity image synthesis with StyleGAN2-ADA[22] and StyleGAN3[23]. Early experiments with conditional architectures such as cSAGAN and cDCGAN? revealed common limitations in low-data, multi-class settings. Despite employing hinge loss and feature matching objectives, these models exhibited mode collapse, class confusion, and blurred textures when trained jointly across six crop–disease categories. Per-class training alleviated some of these issues by reducing distributional complexity, but the resulting images remained inconsistent and insufficiently realistic for adversarial detection research.

To overcome these shortcomings, we transitioned to StyleGAN-based architectures, which have proven reliable for high-resolution synthesis under limited data. StyleGAN2-ADA incorporates adaptive discriminator augmentation to stabilize training when data are scarce, while StyleGAN3 addresses aliasing artifacts and produces smoother interpolations in latent space. For each crop–disease category, we curated approximately 1,000 images from PlantVillage, resized them to 256×256 resolution, and converted them into StyleGAN-compatible formats; forming our input image dataset. Independent StyleGAN2 and StyleGAN3 models were then trained per class in isolated Docker environments, with checkpoints and synthetic samples saved at regular intervals. Training progress is monitored using Fréchet Inception Distance (FID), which provides quantitative insights into model fidelity alongside qualitative evaluation.

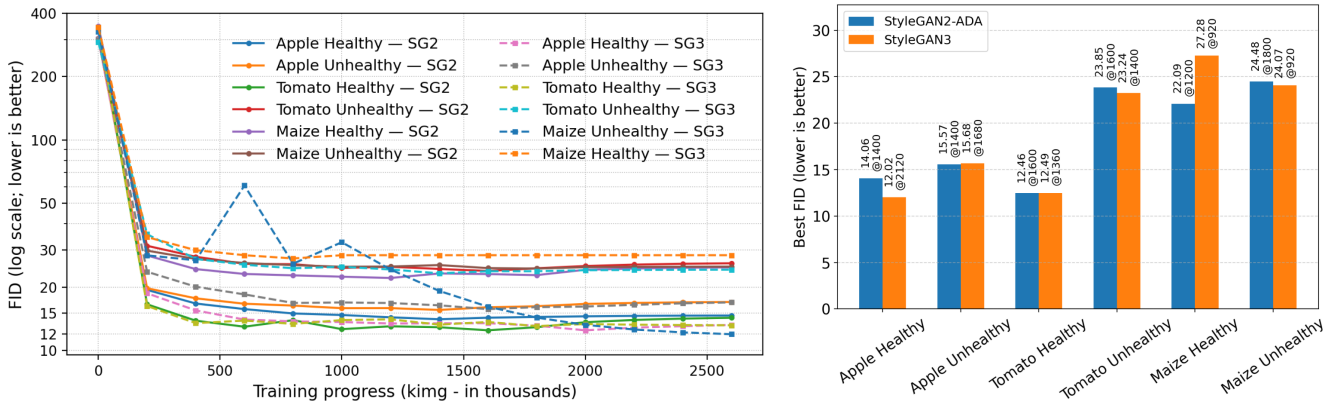


Figure 3. **(a)** Best Fréchet Inception Distance (FID; lower is better) achieved by StyleGAN2-ADA and StyleGAN3 across three crops (Apple, Tomato, Maize), ordered by fruit with *Healthy* first and *Unhealthy* second. Each bar shows the best FID per model–dataset pair, with the training milestone (kimg) annotated above the bar. Overall, StyleGAN3 improves upon or matches StyleGAN2-ADA in several cases (Apple Healthy, Tomato Unhealthy, Maize Unhealthy), is essentially tied on Tomato Healthy, but lags on Apple Unhealthy and Maize Healthy. The optimal training length also varies by dataset and model: for example, Apple Healthy peaks later with StyleGAN3 (2120 kimg) while Maize Unhealthy peaks much earlier (920 kimg). **(b)** Training trajectories of StyleGAN2-ADA and StyleGAN3 across six crop–disease datasets. Each line shows the Fréchet Inception Distance (FID; lower is better) as a function of training progress (in thousands of images, kimg). Healthy and Unhealthy variants for Apple, Tomato, and Maize are included, with StyleGAN2-ADA and StyleGAN3 curves overlaid. This visualization highlights not only the best FID achieved but also the dynamics of convergence, stability, and potential overfitting across datasets.

The qualitative comparison in Table 1 illustrates the relative performance of different generative models alongside original samples for apple, maize, and tomato leaves. Models such as Pix2Pix and DS8 produced recognizable but occasionally distorted images, while BLIP generated diverse samples that often lacked biological plausibility. In contrast, StyleGAN2 and particularly StyleGAN3 yielded highly convincing leaf structures, closely matching the texture and color distributions of original images. These qualitative results are complemented by quantitative analysis, shown in Fig. 3a.

As shown in Fig. 3a, StyleGAN3 often matches or improves upon StyleGAN2-ADA, but the advantage is dataset-dependent. For *Apple (Healthy)*, StyleGAN3 achieves a substantially lower best FID (12.02 vs. 14.06) at a later training stage (2120 vs.

1400 kimg). In contrast, for *Apple (Unhealthy)*, both models perform similarly, with StyleGAN2-ADA slightly ahead (15.57 vs. 15.68). For *Tomato*, StyleGAN3 ties StyleGAN2-ADA on the Healthy set (12.49 vs. 12.46) and outperforms it on the Unhealthy set (23.24 vs. 23.85), converging earlier (1400 vs. 1600 kimg). For *Maize*, StyleGAN2-ADA provides better results on the Healthy set (22.09 vs. 27.28), whereas StyleGAN3 is marginally better on the Unhealthy set (24.07 vs. 24.48), achieving this improvement at a much earlier training point (920 vs. 1800 kimg). These findings indicate that while StyleGAN3 frequently produces sharper details and occasionally converges faster, StyleGAN2-ADA remains competitive and is sometimes superior, particularly for complex datasets such as Maize (Healthy).

In addition to the best FID comparison in Fig. 3a, we further analyze the training dynamics in Fig. 3b, which plots FID trajectories for all six datasets. Several patterns emerge. First, StyleGAN3 generally exhibits smoother convergence with fewer oscillations than StyleGAN2-ADA, especially on *Apple (Healthy)* and *Tomato (Healthy)*, where the FID consistently decreases to the 12–13 range. Second, while both models converge to similar minima on *Tomato (Healthy)*, StyleGAN3 demonstrates a faster drop in FID during early training for *Apple (Healthy)* and *Tomato (Unhealthy)*, suggesting more efficient use of limited data. Third, StyleGAN2-ADA retains an advantage on *Maize (Healthy)*, where its FID stabilizes near 22 compared to StyleGAN3's higher value of 27, indicating that StyleGAN3 may struggle with more complex visual variability in maize leaves. Conversely, StyleGAN3 achieves marginally better results on *Maize (Unhealthy)*, converging earlier and more stably. Together, these observations highlight a nuanced trade-off: StyleGAN3 provides sharper, more stable convergence on certain datasets, whereas StyleGAN2-ADA remains competitive and occasionally superior, particularly for maize leaves with greater intra-class variation.

2.2 Diffuser based Image-to-Image Generation

Each image in our input dataset serves as the starting point for three distinct generative pipelines, all of which receive the same input images and accompanying text prompts. All generation workflows were implemented in Python and executed within Jupyter notebooks using CUDA-enabled GPUs to accelerate inference. This setup allowed for efficient experimentation with model parameters and real-time visualization of outputs.

The three diffusion models evaluated are: (A) Pix2Pix[24], a GPT-3 and Stable Diffusion–driven image-to-image model that performs prompt-based semantic edits without requiring fine-tuning. It was selected for its popularity & ease of use as a purpose built image-to-image model. (B) BLIP[26] is a unified pre-trained vision-language model that supports both image understanding and generation. It improves supervision by bootstrapping captions using synthetic generation and filtering to better align image-text pairs, making it well-suited for image-to-image tasks. BLIPs unique approach, training & ease of use demonstrated reason for inclusion in the study. (C) is an ensemble method combining OpenCV-based segmentation with Dreamshaper-8 inpainting, a fine-tuned stable-diffusion-based in-painting model focusing on making additions to input images according to prompt inputs. The ensemble approach & models selected demonstrate a sophisticated diffusion-based attack focusing on modifications to specific regions of images.

Each workflow was designed to preserve the semantic content of the original image while introducing controlled variations in texture, color, or structure. Each process's output image was evaluated qualitatively through visual inspection of realism and fidelity. The hyper-parameters for each were modified until the output images contained minimal hallucinations, leftover noise, or undesirable deviations from the original input image 5. Detailed information on diffusion processes & hyperparameters is available in the appendix.

2.3 Foundational and Classification Models

We employed three different deep learning architectures to tackle the classification tasks:

- **EfficientNet-B0:** A highly efficient CNN that balances accuracy and computational cost[52].
- **ResNet-50:** A widely-used deep residual network, known for its strong performance on various computer vision tasks[53].
- **CLIP (ViT-B/32):** A vision transformer-based model pre-trained on a large dataset of images and text, which we adapt for our classification task by adding a classification head on top of the frozen visual encoder[54,55].

All models were initialized with pre-trained weights from ImageNet (for EfficientNet and ResNet) or the original CLIP training (for CLIP) to leverage transfer learning [56].

2.4 Experiments

The core of our study involves training the models on three distinct classification schemes to assess their capabilities at different levels of granularity.

1. **Binary Classification:** A simple task to distinguish between *Healthy* and *Unhealthy* leaves, regardless of their origin (original or synthetic).

2. **Generation Source Classification:** A 3-way classification task to identify the origin of an image as *Real*, *GAN-generated*, or *Diffusion-generated*.

3. **Detailed Classification:** The most granular task, a 10-way classification that identifies both the health status and the specific generation source (e.g., "Apple-Healthy-Real", "Apple-Unhealthy-StyleGAN2", etc.).

2.4.1 Training Details

The models were trained for 50 epochs using the Adam optimizer and a Cosine Annealing learning rate scheduler. The initial learning rate was set to 0.001 for the CNN models and a lower rate of 0.0001 for the CLIP-based model to facilitate fine-tuning. We used a weighted Cross-Entropy Loss function to handle potential class imbalances. The dataset was split into 70% for training, 15% for validation, and 15% for testing. Data augmentation techniques, including random flips, rotations, and color jitter, were applied to the training set to improve model generalization.

2.4.2 Evaluation Metrics

Model performance was evaluated using a standard set of metrics: Accuracy, Precision, Recall, and F1-Score (weighted). We also generated confusion matrices to visualize the classification performance for each class.

3 Experimental Results

Our experiments yielded comprehensive results across the three classification schemes for all plant types. The performance of each model was systematically evaluated, and the key findings are summarized in Table 2.

Table 2. Model Performance Across Plants and Classification Types

Plant	Classification	Model	Accuracy	F1 Score	Precision	Recall	Loss	#Samples
Apple	Binary	CLIP	0.9869	0.9869	0.9869	0.9869	0.0369	4500
		ResNet-50	0.9947	0.9947	0.9947	0.9947	**0.0104**	4500
		EfficientNet-B0	**0.9953**	**0.9953**	**0.9948**	**0.9958**	0.0157	4070
	Generation	CLIP	0.9338	0.9413	0.9592	0.9338	0.1851	4500
		ResNet-50	0.9993	0.9993	0.9993	0.9993	0.0058	4500
		EfficientNet-B0	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**0.0001**	4500
	Detailed	CLIP	0.7580	0.7512	0.8054	0.7580	0.4829	4500
		ResNet-50	0.9936	0.9936	0.9938	0.9936	0.0412	4500
		EfficientNet-B0	**0.9969**	**0.9969**	**0.9970**	**0.9969**	**0.0220**	4500
Maize	Binary	**CLIP**	**0.9978**	**0.9978**	**0.9978**	**0.9978**	**0.0034**	2771
		ResNet-50	0.9975	0.9975	0.9975	0.9975	0.0042	2771
		EfficientNet-B0	0.9975	0.9975	0.9975	0.9975	0.0035	2771
	Generation	CLIP	0.8549	0.8645	0.8875	0.8549	0.4255	2771
		ResNet-50	0.9697	0.9704	0.9757	0.9697	0.1251	2771
		EfficientNet-B0	**0.9978**	**0.9978**	**0.9979**	**0.9978**	**0.0147**	2771
	Detailed	CLIP	0.7571	0.7843	0.8638	0.7571	0.6317	2771
		ResNet-50	0.9109	0.9235	0.9446	0.9109	0.2622	2771
		EfficientNet-B0	**0.9971**	**0.9971**	**0.9972**	**0.9971**	**0.0146**	2771
Tomato	Binary	CLIP	0.9975	0.9975	0.9975	0.9975	0.0080	4070
		ResNet-50	0.9978	0.9889	0.9881	0.9898	0.0068	4070
		EfficientNet-B0	**0.9985**	**0.9985**	**0.9985**	**0.9985**	**0.0027**	4070
	Generation	CLIP	0.9469	0.9504	0.9604	0.9469	0.2136	4070
		ResNet-50	0.9978	0.9978	0.9978	0.9978	0.0178	4070
		EfficientNet-B0	**0.9993**	**0.9993**	**0.9993**	**0.9993**	**0.0065**	4070
	Detailed	CLIP	0.8381	0.8399	0.8509	0.8381	0.3967	4070
		ResNet-50	0.9936	0.9936	0.9937	0.9936	0.0368	4070
		EfficientNet-B0	**0.9980**	**0.9980**	**0.9981**	**0.9980**	**0.0143**	4070

3.1 Binary Classification (Healthy vs. Unhealthy)

In the binary classification task, all models demonstrated excellent performance, achieving F1-scores consistently above 0.98. Most of the few errors occurred in cases where unhealthy leaves were misclassified as healthy, particularly in maize. This suggests that certain disease symptoms are visually subtle and can be mistaken for noise or normal texture variation, even by robust models. This indicates that the models can reliably determine the health status of a leaf, even when presented with a mix of original and synthetic images. The performance of EfficientNet-B0, ResNet-50, and CLIP were largely comparable in this task, suggesting that high-level semantic features (health status) are readily learned by all architectures. The accuracy for this task is shown in Table 2.

3.2 Generation Source Classification (Original vs. GAN vs. Diffusion)

When tasked with identifying the source of the images, significant performance differences emerged between the models, as shown in Table 2. The confusion matrices indicate that CLIP frequently confused diffusion-generated maize images with original samples, reflecting its weaker sensitivity to fine-texture cues. ResNet also misclassified a subset of maize GAN images as real, whereas EfficientNet showed minimal overlap, highlighting its superior ability to separate source classes.

- **CLIP** struggled significantly with this task compared to the CNNs. Its F1-scores ranged from 0.658 to 0.751, suggesting that the features learned by CLIP for semantic understanding are less effective at capturing the subtle, low-level artifacts that differentiate original images from synthetic ones.
- **ResNet-50** also performed strongly, with F1-scores above 0.98 for Apple and Tomato. Its performance on the Maize dataset was lower (0.875), indicating some difficulty in distinguishing generative artifacts for that specific plant.
- **EfficientNet-B0** was the standout performer, achieving near-perfect F1-scores across all three plants. It flawlessly identified all images for the Apple dataset and achieved scores of 0.990 and 0.997 for Maize and Tomato, respectively.

3.3 Detailed Classification

The detailed, 10-way classification task proved to be the most challenging, yet it provided the clearest insights into the models' capabilities, as seen in Table 2. The confusion matrices reveal that CLIP's errors were concentrated in maize classes, where original and GAN-generated leaves were often conflated. ResNet exhibited smaller clusters of misclassifications between diffusion- and GAN-generated images, whereas EfficientNet maintained a nearly clean diagonal, demonstrating its robustness even in the most fine-grained attribution setting.

- **CLIP**'s performance was highly variable. While it performed well on Apple (0.891) and Tomato (0.977), its F1-score on the Maize dataset dropped to 0.511. This highlights a lack of robustness in CLIP's ability to handle fine-grained classification of synthetic images, especially for certain data distributions.
- **ResNet-50** also performed well, particularly on the Tomato dataset where it achieved a perfect score. However, its performance on Apple (0.966) and Maize (0.889) was slightly lower than EfficientNet-B0.
- **EfficientNet-B0** once again demonstrated superior performance, achieving perfect or near-perfect F1-scores for all plants. This remarkable result shows that it can learn to simultaneously identify the plant's health status and the specific origin of the image with extremely high precision.

3.4 Analysis

The results consistently show that modern CNN architectures, particularly EfficientNet-B0, are exceptionally effective at detecting synthetically generated images and even identifying their source. A closer comparison across tasks reveals clear performance trends. EfficientNet-B0 remained stable regardless of class granularity, maintaining near-perfect scores from binary to 10-way classification. This indicates that its compound-scaled architecture and supervised meta-learning on ImageNet provide strong inductive priors for handling increasingly complex label spaces. ResNet-50, while generally robust, showed a notable dip in performance on the maize dataset, particularly in the generation classification task (F1 = 0.875). This suggests that ResNet may be less efficient than EfficientNet at capturing the subtle artifacts present in maize images, potentially due to differences in how residual connections emphasize hierarchical features. By contrast, CLIP's performance collapsed as the number of output classes increased: although it was competitive on binary health classification, its F1-score fell as low as 0.511 on the maize detailed classification task. This drop highlights CLIP's sensitivity to class count, reflecting the semantic orientation of its vision–language pretraining. Because CLIP's feature space prioritizes high-level semantic similarity (a "leaf is a leaf"), it struggles to separate fine-grained classes where the differences lie in low-level textures and generative fingerprints. The ability of these models to succeed at the detailed classification task indicates that they are not just learning simple artifacts but are capturing unique feature distributions associated with each generative model.

Conversely, the CLIP model, despite its powerful semantic capabilities, is less adept at this fine-grained detection task. This suggests a fundamental difference in the feature representations of CNNs and Vision Transformers, with CNNs being more sensitive to the kind of low-level texture and frequency artifacts that characterize synthetic images[57]. Our findings underscore the potential of using specialized CNNs as a robust defense mechanism against the injection of synthetic data in agricultural cyber-physical systems.

4 Discussion

4.1 Implications for AI Forensics

Our experiments shed light on the comparative strengths and weaknesses of different AI architectures in the context of synthetic image detection and attribution. CNN-based models, particularly EfficientNet-B0, excelled in distinguishing original from synthetic plant images and in attributing them to their source generative models. This can be explained by their inductive biases: local receptive fields, translation equivariance, and sensitivity to texture equip CNNs to detect the subtle "fingerprints" left by GANs and diffusion models. EfficientNet-B0's compound scaling further enhanced feature extraction efficiency, yielding robust performance even in the most granular 10-way classification task. ResNet-50 also performed strongly, leveraging its residual learning framework to capture discriminative features.

By contrast, CLIP (ViT-B/32) underperformed, especially for maize datasets. Vision transformers lack the strong local inductive biases of CNNs, relying instead on patch-based representations and global self-attention. Combined with CLIP's contrastive pretraining on image–text pairs, this makes them attuned to semantic shape and context rather than fine-grained texture artifacts. While this robustness is advantageous in general vision tasks, it undermines performance in forensic detection tasks requiring sensitivity to pixel-level cues. Moreover, CLIP's performance degraded more sharply than CNNs as class granularity increased, further highlighting the limitations of transformer-based vision–language models for forensic attribution without extensive fine-tuning.

Overall, these findings suggest that CNN-based models pretrained on large-scale image classification are inherently more suited for forensic detection tasks in agriculture, acting as reliable gatekeepers for multi-class attribution frameworks. Transformer-based approaches may still hold promise, but they require significant adaptation to capture subtle generative artifacts.

4.2 Implications for Agricultural Cyber-Biosecurity

From the perspective of agricultural cyber-biosecurity, the implications are critical. High-fidelity GANs and diffusion models can generate synthetic plant images that are visually indistinguishable from original images, creating risks of data poisoning and adversarial manipulation. Adversaries could exploit such models to inject seemingly healthy leaf images that conceal outbreaks, or to fabricate diseased samples that simulate false epidemics. Traditional plant health classifiers—though accurate in distinguishing healthy vs. unhealthy are insufficient for safeguarding agricultural decision pipelines, as they remain vulnerable to such content-based attacks.

Our results confirm that CNN-based detectors can successfully discriminate between original and synthetic data and even attribute images to specific generative sources. This ability is essential for protecting agricultural CPS, ensuring that adversarially generated data does not compromise critical decision-making in disease monitoring, yield prediction, or resource allocation. Importantly, we also observed that CNNs maintained stable performance even as the number of classification categories increased from binary to 10-way attribution. This robustness indicates that these models are scalable and capable of supporting operational pipelines where multiple types of synthetic content may arise simultaneously.

In contrast, models such as CLIP, while powerful in open-domain semantic tasks, currently lack the sensitivity required for agricultural cyber-biosecurity applications. This reinforces the need for agriculture-specific forensics pipelines that prioritize texture and artifact-level detection over general semantic understanding.

Taken together, these findings highlight the necessity of integrating artifact-sensitive AI models into agricultural CPS pipelines. By detecting and attributing synthetic images in real time, CNN-based systems can play a pivotal role in safeguarding food security against emerging generative threats.

4.3 Producer Benefits and Real-World Implications

Producers can benefit from this framework because it improves the reliability of the imagery used to make farm decisions. Detecting and attributing synthetic images prevents falsified data from corrupting disease maps, irrigation schedules, and spray prescriptions. This reduces unnecessary pesticide spraying, overwatering, or excess fertilizer application. It also reduces costs, prevents crop loss, and provides reliable records for insurance, audits, and sustainability reporting.

Real-world systems illustrate these benefits clearly. Autonomous laser weeders rely on cameras to distinguish crops from weeds. If adversarial images suppress weed signals or insert false ones, robots could leave weeds untreated or damage valuable crops. Running authenticity checks inside the vision pipeline reduces this risk and allows robots to switch to a safe mode

when manipulation is detected. Agricultural drones face a similar challenge. UAVs collect images to build NDVI maps and guide variable-rate spraying. A forged tile in the mosaic could mislead farmers into overspraying healthy zones or ignoring real outbreaks. Integrating our detector during image ingestion or map assembly filters out compromised inputs before they drive costly or harmful actions.

The risks also extend to core resource management. Forged images that hide leaf stress could delay irrigation, reducing yields. Synthetic 'stressed' images could trigger overwatering, increase energy use, and promote disease. Misleading nutrient maps could direct fertilizers where they are not needed, creating runoff and environmental damage. By catching these manipulations before they influence operations, our framework safeguards yield, reduces waste, and protects trust in data-driven farming systems.

5 Conclusion

This study evaluated the performance of three state-of-the-art deep learning architectures EfficientNet-B0, ResNet-50, and CLIP (ViT-B/32) in detecting and attributing synthetic plant images generated by GANs and diffusion models. Across binary, generation source, and detailed classification tasks, we observed a clear hierarchy in performance. CNN-based models, particularly EfficientNet-B0, achieved near-perfect accuracy, demonstrating their capacity to capture the fine-grained artifacts left by generative processes. ResNet-50 performed comparably well, though with slightly less consistency across plant types. CLIP, while powerful for semantic tasks, struggled with the low-level distinctions required for this forensic classification[54,58].

These results carry important implications for Agriculture 4.0. As adversaries increasingly leverage generative models to manipulate agricultural data, robust preventative frameworks must be established. CNN-based classifiers, with their architectural advantages and supervised pretraining on natural images, are strong candidates to serve as authenticity filters for agricultural sensing systems. Their deployment can help secure data integrity, enabling reliable disease detection, yield prediction, and resource management in smart farming contexts.

Looking forward, future work should investigate how these models generalize to emerging generative techniques and whether multimodal approaches integrating image, text, and sensor metadata can further improve robustness. The ongoing arms race between generative and discriminative models necessitates proactive research to ensure that agricultural systems remain secure against increasingly sophisticated adversarial threats.

5.1 Future Works

This study opens several promising directions for future research. First, the models should be evaluated against emerging generative techniques, including next-generation diffusion transformers and adversarially trained GANs, to test their resilience under more sophisticated synthetic image distributions. Extending evaluation beyond GANs and diffusion will provide a fuller understanding of the evolving generative landscape.

Second, multimodal approaches warrant exploration. Integrating leaf imagery with text-based descriptions, sensor metadata, or spectral information could improve robustness by providing cross-modal signals for forgery detection. Such approaches may mitigate the limitations observed with CLIP, which struggled with fine-grained artifact detection but excelled at semantic alignment.

Third, scaling to real-world agricultural CPSs requires operational testing. This includes investigating how CNN-based authenticity filters can be embedded within smart farming pipelines for disease monitoring, yield prediction, and resource management, while ensuring computational efficiency on edge devices. Attention should also be given to adversarial scenarios where generative models are used maliciously to conceal or fabricate disease signals.

Overall, future work should move beyond controlled datasets to domain-integrated evaluations, ensuring that synthetic image forensics directly supports the security and reliability of Agriculture 4.0.

References

1. Roopaei, M., Rad, P. & Choo, K.-K. R. Cloud of things in smart agriculture: Intelligent irrigation monitoring by thermal imaging. *IEEE Cloud Comput.* **4**, 10–15, DOI: 10.1109/MCC.2017.18 (2017).
2. Karlov, A. A. Cybersecurity of internet of things—risks and opportunities. In *Proceedings of the XXVI International Symposium on Nuclear Electronics & Computing (NEC'2017)*, 182–187 (Budva, Montenegro, 2017).
3. Araújo, S. O., Peres, R. S., Barata, J., Lidon, F. & Ramalho, J. Characterising the agriculture 4.0 landscape—emerging trends, challenges and opportunities. *Agronomy* **11**, 667 (2021).
4. Malavade, V. N. & Akulwar, P. K. Role of iot in agriculture. *IOSR J. Comput. Eng.* 56–57 (2016).
5. Prasad, R. & Rohokale, V. *Cyber Security: The Lifeline of Information and Communication Technology* (Springer International Publishing, Cham, Switzerland, 2020).

6. Calicioglu, O., Flammini, A., Bracco, S., Bellú, L. & Sims, R. The future challenges of food and agriculture: An integrated analysis of trends and solutions. *Sustainability* **11**, 222, DOI: 10.3390/su11010222 (2019).
7. Devendra, C. *Climate Change Threats and Effects: Challenges for Agriculture and Food Security*. ASM Series on Climate Change (Academy of Sciences Malaysia, Kuala Lumpur, Malaysia, 2012).
8. O'Brien, D. The a to z of cyber security. https://medium.com/threat-intel/the-a-to-z-of-cyber-security-93150c4f336c (2020). Accessed: 2020-09-07.
9. Ivanov, I. Cyber security and cyber threats: Eagle vs "new wars"? https://www.academia.edu/40659366 (2019). Accessed: 2024-07-22.
10. Qazi, S., Khawaja, B. A. & Farooq, Q. U. Iot-equipped and ai-enabled next generation smart agriculture: A critical review, current challenges and future trends. *Ieee Access* **10**, 21219–21235 (2022).
11. Dhar, P. Cybersecurity report: 'smart farms' are hackable farms. *IEEE Spectr.* (2021).
12. Brunelli, D., Albanese, A., Acunto, D. & Nardello, M. Energy neutral machine learning based iot device for pest detection in precision agriculture. *IEEE Internet Things Mag.* **2**, 10–13, DOI: 10.1109/IOTM.0001.1900037 (2019).
13. Alsamhi, S. H. *et al.* Drones' edge intelligence over smart environments in b5g: Blockchain and federated learning synergy. *IEEE Transactions on Green Commun. Netw.* **6**, 295–312, DOI: 10.1109/TGCN.2021.3132561 (2022).
14. Paniagua, T., Savadikar, C. & Wu, T. Adversarial perturbations are formed by iteratively learning linear combinations of the right singular vectors of the adversarial jacobian. In *Forty-second International Conference on Machine Learning*.
15. Unal, Z. Smart farming becomes even smarter with deep learning—a bibliographical analysis. *IEEE Access* **8**, 105587–105609, DOI: 10.1109/ACCESS.2020.3000175 (2020).
16. Radoglou-Grammatikis, P., Sarigiannidis, P., Lagkas, T. & Moscholios, I. A compilation of uav applications for precision agriculture. *Comput. Networks* **172**, 107148, DOI: 10.1016/j.comnet.2020.107148 (2020).
17. Lottes, P., Khanna, R., Pfeifer, J., Siegwart, R. & Stachniss, C. Uav-based crop and weed classification for smart farming. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 3024–3031, DOI: 10.1109/ICRA.2017.7989347 (2017).
18. Alferidah, D. K. & Algosaibi, A. The effect of adversarial machine learning attack on agriculture field and food security. In *2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS)*, 1–10, DOI: 10.1109/ICDS62089.2024.10756330 (2024).
19. Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A. & Madry, A. Raising the cost of malicious ai-powered image editing (2023). 2302.06588.
20. Rahman, Z. U., Asaari, M. S. M., Ibrahim, H., Abidin, I. S. Z. & Ishak, M. K. Generative adversarial networks (gans) for image augmentation in farming: A review. *IEEE Access* **12**, 179912–179943, DOI: 10.1109/ACCESS.2024.3505989 (2024).
21. Ghazal, S., Munir, A. & Qureshi, W. S. Computer vision in smart agriculture and precision farming: Techniques and applications. *Artif. Intell. Agric.* **13**, 64–83, DOI: https://doi.org/10.1016/j.aiia.2024.06.004 (2024).
22. Karras, T. *et al.* Analyzing and improving the image quality of stylegan. *Proc. CVPR* (2020).
23. Karras, T. *et al.* Alias-free generative adversarial networks. *Proc. NeurIPS* (2021).
24. Brooks, T., Holynski, A. & Efros, A. A. Instructpix2pix: Learning to follow image editing instructions (2023). 2211.09800.
25. Rombach, R. *et al.* High-resolution image synthesis with latent diffusion models. *Proc. CVPR* (2022).
26. Li, J., Li, D., Xiong, C. & Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML* (2022).
27. Yaseen, M. U. & Long, J. M. Laser weeding technology in cropping systems: A comprehensive review. *Agronomy* **14**, DOI: 10.3390/agronomy14102253 (2024).
28. Zhang, Y., Li, Y., Li, Y. & Guo, Z. A review of adversarial attacks in computer vision (2023). 2308.07673.
29. Goodfellow, I. *et al.* Generative adversarial networks. In *Advances in neural information processing systems* (2014).
30. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, 2256–2265 (PMLR, 2015). ISSN: 1938-7228.
31. Sarala, P., Siripurapu, P., Chandrika, V. L., Prattipati, D. & Manoj, G. S. Deepfake detection using deep learning. *Int. J. on Sci. Technol.* **16** (2025).

32. Afchar, D. *et al.* Mesonet: a compact facial video forgery detection network. *Proc. WIFS* (2018).

33. Tolosana, R. *et al.* Deepfakes evolution: Analysis of facial manipulation videos. In *Proc. IEEE SMC* (2020).

34. Cai, A. *et al.* Frequency-aware feature fusion for cross-model synthetic image detection. *J. Digit. Forensics* (2024).

35. Guo, A. *et al.* Multi-branch rgb and frequency-domain feature extraction for synthetic image detection. *IEEE Transactions on Inf. Forensics Secur.* (2023).

36. Chang, Y.-M., Yeh, C., Chiu, W.-C. & Yu, N. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419* (2023).

37. Sha, A. *et al.* De-fake: Detection of ai-generated images using multimodal cues. In *Proc. ICCV Workshops* (2022).

38. Wang, S.-Y. *et al.* Cnn-generated images are surprisingly easy to spot... for now. *Proc. CVPR* (2020).

39. Luo, Z., Li, Q. & Zheng, J. A study of adversarial attacks and detection on deep learning-based plant disease identification. *Appl. Sci.* (2021).

40. ling You, H., Lu, Y. & Tang, H. Plant disease classification and adversarial attack using simam-efficientnet and gp-mi-fgsm. *Sustainability* (2023).

41. Li, Y. & Lu, Y. Plant disease classification and adversarial attack based cl-condensenetv2 and wt-mi-fgsm. *Int. J. Adv. Comput. Sci. Appl.* (2023).

42. Echim, S.-V., Taiatu, I.-M., Cercel, D.-C. & Pop, F.-C. Explainability-driven leaf disease classification using adversarial training and knowledge distillation. In *International Conference on Agents and Artificial Intelligence* (2023).

43. Yang, W. *et al.* Adversarial training collaborating multi-path context feature aggregation network for maize disease density prediction. *Processes* (2023).

44. Nazki, H., Yoon, S., Fuentes, A. & Park, D. Unsupervised image translation using adversarial networks for improved plant disease recognition. *Comput. Electron. Agric.* **168** (2019).

45. Bi, L. & Hu, G. Improving image-based plant disease classification with generative adversarial network under limited training set. *Front. Plant Sci.* **11** (2020).

46. Wang, D. *et al.* Early detection of tomato spotted wilt virus by hyperspectral imaging and outlier removal auxiliary classifier gans. *Sci. Reports* **9** (2019).

47. Alshammari, K., Alshammari, R., Alshammari, A. & Alkhudaydi, T. An improved pear disease classification approach using cycle generative adversarial network. *Sci. Reports* **14** (2024).

48. Singh, A. K., Rao, A., Chattopadhyay, P., Maurya, R. & Singh, L. Effective plant disease diagnosis using vision transformer trained with leafy-gan-generated images. *Expert. Syst. with Appl.* **254**, 124387 (2024).

49. Chen, J. *et al.* Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis Mach. Intell.* **47**, 961–977, DOI: 10.1109/TPAMI.2024.3480519 (2025).

50. Yu, N., Davis, L. & Fritz, M. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7556–7566 (2019).

51. Marra, F., Gragnaniello, D., Verdoliva, L. & Poggi, G. Do gans leave artificial fingerprints? *Signal Process. Image Commun.* **74**, 65–75 (2019).

52. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning* (2019).

53. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016).

54. Radford, A. *et al.* Learning transferable visual models from natural language supervision (2021). 2103.00020.

55. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale (2021). 2010.11929.

56. Kornblith, S., Shlens, J. & Le, Q. V. Do better imagenet models transfer better? (2019). 1805.08974.

57. Naseer, M. *et al.* Intriguing properties of vision transformers (2021). 2105.10497.

58. Xie, C. *et al.* Fg-clip: Fine-grained visual and textual alignment (2025). 2505.05071.

59. Li, D., Li, J. & Hoi, S. C. H. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing (2023). 2305.14720.

A Appendix: Diffusion model processes, prompts, & hyperparameter configurations

A.1 Prompt Design

A consistent natural language prompt was used across all diffusion-based models in this study, including Pix2Pix, BLIP, and Diffusion-DS8. This prompt was carefully crafted to ensure that each model performed localized edits focused solely on the leaf element, while preserving the overall integrity and realism of the original image. The prompt was as follows:

> *"Generate a high-fidelity image closely resembling the input, preserving key details such as composition, color balance, and lighting. Maintain a realistic style, ensuring textures, depth, and fine details remain natural and coherent. Modify only the leaf element in the image, keeping the background, surrounding objects, and lighting conditions completely unchanged. Do not alter the context, perspective, or environment outside of the leaf. Ensure the leaf's integration remains consistent with the original image's style and realism."*

This prompt was applied uniformly across input images to guide the generation process in a controlled and consistent manner. Its design aimed to minimize hallucinations and ensure that the synthetic outputs remained visually and semantically aligned with the original inputs, regardless of the underlying diffusion model used.

Dataset and Workflow Each image was processed individually using each of the diffusion-based models evaluated in this study, including Pix2Pix, BLIP, and Dreamshaper-8. This resulted in a one-to-one mapping between input and output images for each model, enabling direct comparison of their generative performance across identical input conditions.

A.2 Instruct pix2pix

To generate high-fidelity synthetic images with localized edits, we employed the **Pix2Pix** model developed by Brooks et al.[24], available via Hugging Face at `https://huggingface.co/timbrooks/instruct-pix2pix`. This diffusion-based model enables fine-grained image manipulation guided by natural language instructions, making it well-suited for tasks requiring subtle, localized changes while preserving global image fidelity.

Hyperparameter Configuration After iterative experimentation using the Pix2Pix Pipeline from Hugging Face's diffusers library, the following hyperparameters were selected to enforce strict adherence to both the input image and the textual prompt:

- **CFG Scale:** 25
- **Strength:** 0.05
- **Inference Steps:** 50

A high Classifier-Free Guidance (CFG) scale of 25 was used to strongly condition the model on both the prompt and the input image, ensuring that the generated outputs closely followed the intended transformation. The low strength value of 0.05 limited the degree of deviation from the original image, preserving its structure while allowing for subtle, instruction-driven edits. A high number of inference steps (50) provided ample opportunity for the diffusion process to refine the output, resulting in visually coherent and high-fidelity images across batch processing.

A.3 Salesforce-BLIP Diffusion

We also utilized the **Salesforce BLIP-Diffusion** model, available via Hugging Face[59], to perform concept-conditioned image generation. This model enables targeted editing by conditioning on a visual concept (e.g., "leaf") alongside a natural language prompt, making it suitable for localized perturbations with high perceptual fidelity.

Hyperparameter Configuration The following hyperparameters were applied using the `BlipDiffusionPipeline` from Hugging Face's diffusers library:

- **Guidance Scale:** 7.5
- **Inference Steps:** 50
- **Resolution:** 512×512 pixels

Both the `conditional` and `target` subjects were set to `leaf`, prompting the model to focus modifications strictly on the leaf region. A curated negative prompt was used to suppress artifacts such as blur, noise, and anatomical inconsistencies.

Implementation Workflow Images were processed in class-based batches using a directory structure identical to the one used for other diffusion models. For each input, BLIP-Diffusion generated an edited version conditioned on the standardized prompt, and output files were saved with a `_blipdiffusion` suffix. The pipeline was executed on GPU with

`torch.float16` precision for improved efficiency. Error handling and progress monitoring were implemented using standard Python tools.

The BLIP-generated images were visually high quality and semantically close to the originals, providing effective adversarial samples for evaluating classifier robustness.

A.4 OpenCV & Dreamshaper8-inpainting

To explore the potential of adversarial image manipulation in a computer vision context, a custom pipeline was developed using OpenCV for segmentation and the DreamShaper 8 inpainting model via Hugging Face's diffusers library. The primary objective was to replace leaf regions in plant images with visually similar but synthetically generated leaves that could plausibly deceive a downstream classification model. This method also holds potential for simulating both healthy-to-unhealthy and unhealthy-to-healthy transformations, depending on the prompt used.

- **Model:** `Lykon/dreamshaper-8-inpainting`
- **Guidance Scale:** 7.5
- **Inference Steps:** 50
- **Prompt:** *"Replace the masked area with a healthy green leaf that perfectly matches scene lighting, shadows, color tone, perspective, and scale. Leave everything else untouched."*
- **Negative Prompt:** *"person, human, hand, animal, insect, text, logo, tool, vehicle, blur, noise"*

The segmentation process employed a hybrid approach combining HSV thresholding and morphological operations with a guarded GrabCut refinement. This ensured accurate isolation of the leaf region while avoiding over-segmentation. The resulting binary mask was used to guide the inpainting process, which replaced the masked region with a synthetic leaf that matched the surrounding visual context.3

The DreamShaper 8 model was selected for its accessibility, popularity, and effectiveness in generating high-quality inpainted results. The prompt was iteratively refined to ensure realism and consistency across diverse input images. Final outputs were composited with the original background to preserve non-leaf regions, resulting in believable synthetic images suitable for adversarial testing scenarios.

Class	Original Leaf Image	Mask created By OpenCV	Composite created by DS8
Apple			
Tomato			

Table 3. OpenCV & Dreamshaper-8 Process; Apple & Tomato Classes

A.4.1 Maize-Specific Adjustment for DreamShaper 8 Inpainting

To address limitations observed in the original DreamShaper 8 inpainting pipeline when applied to close-up images of maize leaves, a targeted adjustment was developed. The original segmentation method, which relied on global green-hue thresholding

and GrabCut refinement, often failed to isolate meaningful regions in macro photographs of maize foliage. This led to suboptimal inpainting results for this specific class of images.

The revised approach introduces a feature-driven segmentation strategy based on superpixel analysis and local texture variance. Specifically, the image is first decomposed into approximately 200 superpixels using the SLIC (Simple Linear Iterative Clustering) algorithm. Each superpixel is then scored by the variance of its luminance (L channel) in Lab color space, capturing local texture complexity. The top 5 regions with the highest variance are selected as the primary features for replacement. These regions are slightly dilated to ensure full coverage and then merged into a single binary mask for inpainting. 4

The inpainting prompt was also adapted to reflect the close-up nature of the input data:

- **Prompt:** *Macro photograph of a maize leaf—replace the masked region with a plausible, unrecognizable maize-leaf surface, matching color, lighting and scale perfectly, no trees, no sky, no people.*

This prompt is designed to ensure that the synthetic samples remained visually consistent with the original image while introducing subtle, realistic alterations. The specificity of the prompt and the localized masking strategy together enabled more effective manipulation of maize leaf textures, particularly in scenarios where the goal is to simulate deceptive modifications for downstream classification tasks.

As with the original DS8 pipeline, the inpainted region is composited back into the original image using the generated mask. The effectiveness of this maize-specific adjustment is evaluated visually, with particular attention paid to the believability of the composite and the fidelity of the mask to prominent leaf features.

Original leaf image | Mask created by OpenCV | Changes in circled regions

Table 4. OpenCV & DS8 Inpainting process; Maize Adjustment

To evaluate the feasibility of detecting synthetically generated agricultural imagery, we designed a comprehensive experimental pipeline. This section details the dataset, the classification models, the experimental setup, and the evaluation metrics used in this study.

B Appendix: Diffusion model outputs before & after hyper-parameter tuning

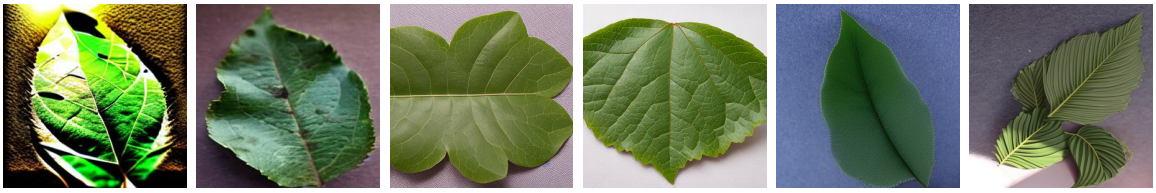
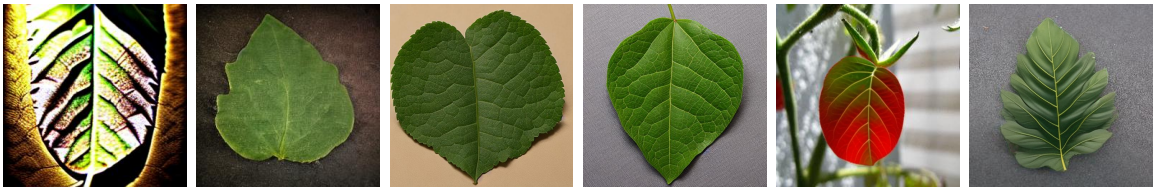
Crop	Instruct Pix2Pix		BLIP		OpenCV & Dreamshaper 8	
	Before	After	Before	After	Before	After
Apple-H						
Apple-Un						
Maize-H						
Maize-Un						
Tomato-H						
Tomato-Un						

Table 5. Comparison of diffusion-generated images exhibiting errors before & resolved after hyperparameter tuning. H = healthy leaf, Un = unhealthy leaf.