# CONCEPT-GUIDED GAN WITH DISENTANGLED REPRESENTATIONS FOR EXPLAINABLE PLANT DISEASE IMAGE SYNTHESIS

***Zahra Shams Khoozani[1], Aznul Qalid Md Sabri[1*], Woo Chaw Seng[1], Manjeevan Seera[2]***

[1] Department of Artificial Intelligence, Faculty of Computer Science & Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia

[2] School of Business, Monash University Malaysia, Selangor, Malaysia

Emails: zahra.shams@um.edu.my[1], aznulqalid@um.edu.my[1*](Corresponding Author), cswoo@um.edu.my[1], manjeevansingh.seera@monash.edu[2]

***ABSTRACT***

*Generative Adversarial Networks (GANs) have advanced image synthesis and are widely used to augment training data for deep Convolutional Neural Networks (CNNs). However, in scientific domains like plant disease identification, interpretability and morphological control are essential. Existing GANs typically rely on pixel-level feedback from the discriminator and function as black-box architectures, often learning entangled representations that limit control over high-level morphological features like leaf shape and disease patterns. This study introduces CXAI-GAN, an explainable GAN that integrates Concept-based Explainable AI (C-XAI) to generate biologically realistic images with explicit morphological control. The generator is modified to disentangle and encode three interpretable concepts: leaf shape, surface texture, and disease pattern. Unlike post-hoc Explainable AI (XAI) methods (e.g., LIME, SHAP) that reveal what features matter numerically, CXAI-GAN explains why outputs are generated through concept learning. CXAI-GAN achieves strong performance with an FID of 19.64, SSIM of 0.955, and PSNR of 34.91. Fine-grained evaluations show high fidelity: shape similarity (HDS 0.973), texture alignment (VPS 0.999), and local SSIM 0.937. In a binary classification task of visually similar grape diseases, CXAI-GAN improved accuracy by 10% and reached 96.7% with synthetic training. These results demonstrate CXAI-GAN's effectiveness in generating interpretable, high-quality images for downstream scientific tasks.*

***Keywords: Explainable AI; Generative adversarial network; Plant disease image synthesis; Disentangled representation learning; Data augmentation.***

## 1.0 INTRODUCTION

Generative adversarial networks (GANs) [1] are a transformative class of convolutional neural networks (CNNs) that have achieved remarkable success in synthetic image generation. The core mechanism of GANs involves an adversarial training process, where a CNN discriminator guides the generator to iteratively improve output quality by distinguishing real from generated samples. This setup enables the production of visually realistic images that closely resemble real data, making GANs powerful tools for tasks such as data augmentation, image-to-image translation, and image registration. However, despite their success, GANs face critical limitations when applied to scientific domains that demand interpretability and morphological fidelity [2].

While state-of-the-art models such as StyleGAN [3], SC-GAN [4], hybrid Diffusion-StyleGAN [5], and xAI-GAN [6] architectures offer significant advances in visual fidelity, they primarily focus on improving visual quality and provide limited control over high-level and biologically meaningful features. StyleGAN [3] introduces a style-based architecture that enables limited, scale-specific control of visual attributes such as position and orientation through an intermediate latent space ($W$). SC-GAN [4] attempts to disentangle semantic components such as style and content to improve structured generation. xAI-GAN [6] incorporates post-hoc explanation methods like LIME and SHAP to enhance interpretability but remains limited to providing post-hoc insights rather than interactive control, falling short in enabling expert-in-the-loop frameworks necessary for scientific applications. Hybrid models that combine StyleGAN with diffusion approaches [5] improve both fidelity and diversity but further complicate interpretability. Despite these advances, the latent spaces of these models often remain entangled, limiting interpretability and control over biologically relevant features.

A critical limitation of GANs lies in their inherent black-box nature and lack of interpretability [7]. These models learn complex, high-dimensional mappings from latent vectors to output images without offering interpretable insights into how specific features are generated in synthetic images. This lack of transparency poses a significant barrier in scientific domains where trust, validation, and precise control are essential. In plant disease identification, for instance, domain experts require confidence that a generated lesion pattern accurately reflects

not only visual realism but also the underlying biological processes of infection. Without mechanisms to understand or influence the internal representations driving generation, GANs remain difficult to validate, interpret, or safely deploy in critical scientific and clinical applications.

Another major drawback in GAN architectures stems from their reliance on pixel-level generation, which emphasizes low-level visual accuracy over high-level and meaningful morphological understanding. These models are typically optimized using pixel-wise loss functions or adversarial feedback that reward visual plausibility yet lack awareness of conceptual or morphological correctness. As a result, GANs may produce images that appear realistic to automated metrics but fail to retain biologically significant features, such as infected leaf shape, venation distortion, or disease patterns. Morphological attributes are critical diagnostic indicators in accurate disease identification, as they often reflect specific disease processes or stages of progression. The pixel-level approach encodes information as numerical arrays, which may be effective for machine learning algorithms but are inherently not interpretable to human experts. Experts interpret data through conceptual reasoning and high-level morphological features such as shape or pigments. Consequently, a significant disconnect emerges between the model's internal representation learning and the expert's domain knowledge, limiting the practical utility of GANs in scientific settings where expert validation, fine-grained morphological control, and interpretability are essential.

A further significant limitation in GANs is the entanglement of latent representations, where multiple biological features are jointly encoded and cannot be independently controlled. In these representations, attempting to modify one feature often leads to unintended changes in others, making it difficult to isolate and control individual attributes. This lack of representational separability presents a critical obstacle in scientific applications, where interpretability, reproducibility, and controlled experimentation are essential. For example, in a disease identification task, a domain expert may desire to adjust the morphology of a disease lesion while preserving the leaf's venation and structural outline. Entangled representations make such targeted interventions infeasible, limiting the model's utility for tasks such as controlled lesion progression analysis or generating synthetic datasets for model training under specific biological constraints.

Moreover, these challenges lead to several broader technical limitations of GANs in scientific applications, as summarized in Table 1. These include the need for large training datasets (often 5,000 - 10,000 images per class), extensive training iterations (2,000 - 5,000 epochs), and substantial computational resources [8] [9] [10]. Training is computationally intensive, often requiring high-performance GPUs or TPU clusters, with runtimes ranging from several hours to multiple days to produce high-quality outputs. Standard loss functions frequently yield overly smooth results that lack fine-grained biological detail, while persistent issues such as mode collapse reduce output diversity and compromise model robustness [8] [11]. Furthermore, in the absence of structured and interpretable latent spaces, validating the biological plausibility of generated samples remains a significant challenge, particularly in domains like plant pathology, where semantic accuracy is critical [12] [13].

Table 1: Limitations of existing GANs in biological image synthesis.

| **GAN Limitations** | **Ref.** | **Specific Impact on Biological Image Synthesis** |
|---|---|---|
| Black-Box Architecture | [12] | Limits interpretability and transparency; difficult to understand how specific biological or morphological features influence generated images; reduces user trust. |
| Pixel-Level Generation | [8] | Lacks semantic understanding of learned features; reduces biological realism; slow convergence, increase training time, and high computational cost. |
| Entangled Latent Representations | [13] | Encodes multiple attributes within a single latent space; limits independent feature control; lacks human-in-the-loop and explainability. |
| Mode Collapse | [11] | Results in highly repetitive or overly similar outputs, reducing the diversity essential for robust disease modelling and generalization. |
| Poor High-Level Correlation | [9] | Fails to preserve meaningful relationships between leaf structure (e.g., venation) and disease regions, reducing biological realism. |
| Weak Structural Consistency | [14] | Can produce distorted outputs with inconsistent structures, impacting practical usability. |

While post-hoc explainability techniques such as LIME (Local Interpretable Model-agnostic Explanations) [15], SHAP (SHapley Additive exPlanations) [16], Grad-CAM [17], and Saliency Maps [18] have been developed to improve model transparency, their effectiveness in biological domains remains limited. These methods typically operate on low-level input gradients or perturbations and are designed to answer what influenced a model's decision, such as identifying which image regions contributed most to a classification. However, they often fail to explain why a particular prediction was made in terms of high-level, domain-relevant concepts. For instance, a saliency map may highlight a lesion region, but it cannot indicate whether the model attended to diagnostic features like lesion shape, margin irregularity, or venation distortion. This semantic gap limits their practical utility in

scientific contexts where concept-level reasoning, domain alignment, and causal interpretability are critical for trust and adoption.

To address this gap, there is a growing need for GAN frameworks that incorporate human-in-the-loop (HIL) architectures [19] [20], enabling domain experts to guide, evaluate, and refine the generative process based on biologically meaningful criteria. Such systems should support interactive control over semantic features and offer transparent evaluation mechanisms to test, for example, how specific morphological variations (e.g., lesion shape or spread) influence downstream classification or diagnostic performance. Embedding disentangled and concept-aligned control into the generative process offers a promising solution. By explicitly modelling distinct conceptual features, such as leaf shape, venation texture, and disease lesion patterns, it can produce interpretable outputs that align more closely with expert knowledge and reasoning.

To address these limitations, we introduce CXAI-GAN, a Concept-based Explainable Generative Adversarial Network designed to overcome critical challenges of GANs in scientific domains, including their reliance on pixel-level generation, their inherent black-box architecture, and the entanglement of learned representations. CXAI-GAN integrates principles from Concept-based Explainable AI (C-XAI), a new research direction within Explainable AI (XAI), into the generative process by learning biologically meaningful and conceptual representations such as leaf shape, textural patterns, and disease features that are interpretable and understandable to domain experts. At the core of the architecture is a Concept-based Explainable and Disentangled (CXD) generator that enables independent control over these high-level attributes, moving beyond low-level pixel constraints. The proposed framework is comprehensively evaluated through both quantitative image quality metrics (e.g., FID, SSIM, PSNR, VPS, HDS) and downstream classification performance, demonstrating its effectiveness in generating plant disease images that are not only biologically accurate but also offer transparent, human-understandable explanations. Fig. 1 illustrates the overall system architecture of the proposed CXAI-GAN. This study addresses three main research questions:

1. How can concept-based explainability be embedded into GAN architectures to support interpretable plant disease image generation?
2. What generator design enables disentangled learning and independent control over key morphological and biological features such as shape, texture, and disease regions?
3. How effective is CXAI-GAN in producing realistic, semantically meaningful images that improve classification performance?

The main contributions are:

1. A new GAN framework, CXAI-GAN, that integrates concept-based explainability to overcome limitations of black-box generation and low-level pixel synthesis.
2. A Concept-based Explainable and Disentangled (CXD) generator that models shape, texture, and disease features in independent latent spaces with structured control.
3. A comprehensive evaluation showing high image fidelity (FID 19.64, SSIM 0.955, PSNR 34.91) and improved classification performance, including an 10% gain in a difficult disease pair.

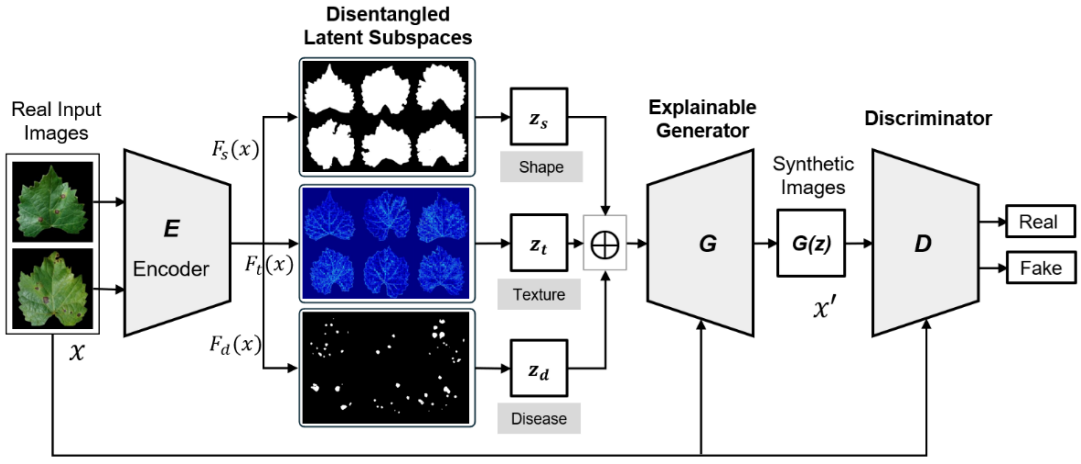


Fig. 1: Overview of the proposed CXAI-GAN system architecture. The encoder disentangles the input into latent subspaces: shape, texture, and disease patterns. These interpretable concepts are then used by the Concept-based Explainable and Disentangled (CXD) generator to synthesize high-fidelity, biologically meaningful images, enabling a concept-based explainable AI (C-XAI) approach.

## 2.0 RELATED WORK

### 2.1 Existing Generative Adversarial Networks (GANs)

The original GAN, introduced by [1], consists of a generator-discriminator pair trained through adversarial learning. While this approach demonstrated the feasibility of data-driven image synthesis, it relied on pixel-level generation from a single latent space, leading to entangled feature representations that limit interpretability. Conditional GANs (cGANs) [21] extended the original GAN formulation by conditioning generation on external class labels or auxiliary information. Although this facilitated some level of control over output characteristics, cGANs still suffered from implicit feature mixing, as they lacked a structured latent space for independent attribute control.

InfoGAN [22] introduced an unsupervised mechanism to encourage disentanglement by maximizing the mutual information between a subset of latent codes and the generated output. While this allowed the discovery of limited interpretable factors such as width and orientation without labels, the disentanglement remained implicit, coarse-grained, and generic, often misaligned with semantically meaningful high-level features, thereby limiting its effectiveness in biologically grounded contexts. Subsequent works, such as Pix2Pix [23] and CycleGAN [24], introduced conditional and unpaired image-to-image translation techniques, respectively. However, these models remained fundamentally black-box architectures without explicit mechanisms to disentangle high-level features or enforce structural consistency.

To address the limitations of pixel-level generation in GANs, recent advances such as the StyleGAN [3] family, including StyleGAN2 [25] and StyleGAN3 [26], introduced style-based architectures that modulate feature maps via intermediate latent spaces and per-layer adaptive style injections. While this design enables limited control over certain image features such as pose (i.e., object orientation or position), global lighting, and overall texture style, it lacks semantic supervision and does not integrate explicit biologically meaningful concepts. As a result, the model fails to learn disentangled or interpretable representations of specific concepts, such as lesion shape or disease texture, and instead encodes entangled visual abstractions that are not directly mappable to meaningful biological attributes.

GANSpace [27] and InterFaceGAN [28] are post-hoc methods developed to improve the interpretability of pretrained StyleGANs by identifying latent directions that correspond to limited semantic variations, such as lighting, orientation, age, gender, or expression. GANSpace applies unsupervised principal component analysis (PCA) to uncover these directions, while InterFaceGAN employs SVM supervised classifiers to identify attribute boundaries for directional manipulation. However, both techniques operate without architectural modifications or concept-level supervision, resulting in discovered directions that often entangle multiple attributes. Therefore, they lack explicit semantic alignment between latent directions and the resulting output attributes, limiting their effectiveness for controlled, concept-aware generation.

Existing disentangled GANs, such as SC-GAN [4], attempt to separate latent factors by encouraging statistical dependency between latent codes and output features through auxiliary objectives, but they lack explicit concept-level supervision. This often results in entangled or unstable representations, especially in tasks requiring fine-grained, feature control. Moreover, their disentanglement is typically limited to abstract visual styles (e.g., pose, appearance) rather than interpretable or biologically meaningful attributes. In SC-GAN, disentanglement is applied primarily at the latent code level, where different segments of the input latent vector are intended to control separate attributes. However, during feature propagation through the generator, entanglement re-emerges in intermediate layers, especially in deeper convolutional layers where global and local features interact without semantic constraints. As a result, the weak semantic alignment makes it difficult to achieve consistent and independent control over conceptual attributes in the generated output.

GANterfactual [29] generates counterfactual examples by perturbing latent representations in pretrained GANs, typically using supervised attribute classifiers. However, it lacks an explicitly structured latent space, resulting in coarse or entangled modifications that may simultaneously affect multiple attributes. This limits its applicability in domains requiring fine-grained semantic control and high-fidelity concept preservation, as the latent changes are not aligned with disentangled representations. xAI-GAN [6] introduces pixel-level interpretability by integrating saliency maps and attribution-guided attention mechanisms into the adversarial training process. While this enhances model transparency, the latent space remains weakly structured and semantically entangled, providing limited support for controllable generation. Furthermore, its interpretability is post-hoc and global, lacking enforcement of localized or concept-specific disentanglement, which restricts its applicability in concept-level generative frameworks.

ControlNet [30] augments diffusion-based or conditional generative models by injecting external structural priors (e.g., edges, depth, pose) into specific layers of the generation process, enabling strong spatial conditioning. However, ControlNet is primarily designed for pixel-level structural control and lacks mechanisms for high-level conceptual disentanglement. Therefore, it is limited in its effectiveness for explainable or scientific image synthesis tasks. Diffusion-StyleGAN hybrids [5] extend these models by incorporating iterative denoising to improve sample fidelity, yet they inherit the same interpretability constraints, further exacerbated by the stochastic nature of the diffusion process. Consequently, both StyleGAN variants and their diffusion-based extensions remain

fundamentally black-box models, limiting their use in applications requiring explicit, biologically interpretable control, such as plant disease image synthesis.

In summary, existing GANs face two key drawbacks that reduce their effectiveness in scientific image synthesis, as summarized in Table 2. Some, like StyleGAN or InfoGAN, introduce partial latent disentanglement to control limited abstract attributes such as lighting or orientation but, lack interpretable, concept-level control over biologically meaningful morphological features. As a result, partial disentanglement restricts the independent learning and control of biological features such as leaf shape or disease patterns. Others adopt post-hoc explainability methods, such as xAI-GAN with LIME or SHAP, which interpret model decisions after training. However, these methods retain pixel-level generative processes with entangled latent representations and black-box architectures, offering limited control and producing explanations disconnected from the model's internal learning process. Consequently, current GANs do not offer a unified framework that combines human-understandable explainability with explicit and biologically relevant feature control. In contrast, this study proposes a concept-based explainable GAN with disentangled feature learning, enabling explicit control over morphological features. This facilitates interpretable and controllable generation aligned with expert needs and supports human-in-the-loop integration for domain-guided synthesis.

### 2.2 Concept-Based Explainable AI (C-XAI) in GANs

Explainability is critical in modern machine learning, particularly in scientific and high-stakes domains where transparency and trust are essential. While GANs have demonstrated impressive capabilities in synthesizing realistic data, they remain largely opaque, limiting their applicability in domains that require interpretability. Traditional post-hoc explainability methods such as LIME, SHAP, and Grad-CAM, which attribute model decisions to input features, are primarily designed for discriminative models. These pixel-level attribution techniques offer limited insight into the generative process, as they are global, non-causal, and disconnected from the model's representation learning. Consequently, they are insufficient for understanding how GANs manipulate internal representations or for enabling precise user control over the generated output.

Recent advances have shifted toward concept-based explainability [31], which aims to interpret model behaviour in terms of high-level semantic features that align closely with human conceptual reasoning. The term 'concept' refers to high-level features (such as shape, texture, or disease characteristics) that reflect how humans naturally interpret and categorize information. These concepts are meaningful to experts and remain interpretable across different instances. Rather than attributing outputs to individual pixels, concept-based approaches disentangle the latent space into modular components, each corresponding to a distinct and controllable concept. This framework offers several key advantages. It enables modular, localized interpretation and supports human-in-the-loop interaction, allowing domain experts to guide or validate concept manipulations during training and inference. Additionally, it facilitates causal reasoning by modelling how changes in each concept affect the output and improves model accountability in scientific contexts where explanations must be grounded in expert knowledge. Thus, concept-based explainability advances beyond traditional attribution by offering a structured and interpretable interface between models and users. Its alignment with human reasoning, support for expert interaction, and capacity for concept-level control make it particularly suitable for domains that demand transparency and scientific evaluation.

Several concept-based explainable models have emerged in recent years, aiming to bridge the gap between black-box model predictions and human reasoning. Notable examples include the Concept Bottleneck Model (CBM) [32] and ConceptTransformer [33], which enforce the explicit prediction of human-defined concepts prior to final decision-making. TCAV [34] measures a model's sensitivity to user-defined concepts by analyzing internal activations. Other approaches, such as ACE [35], ICE [36], and CAR [37], are unsupervised methods that automatically discover meaningful concepts without relying on labeled data. These models have been primarily applied to downstream tasks such as classification and segmentation, where concept-level interpretability aids in understanding and validating predictions. However, their application to generative models like GANs remains limited. Existing methods, such as GANSpace and InterFaceGAN, attempt to link latent directions to semantic variations, but they operate on black-box GAN architectures without explicit concept supervision or structural transparency.

As a result, these methods rely on pixel-level and post-hoc interpretability techniques that do not support morphologically meaningful or causal relationships between the learning process and output attributes. Although approaches like xAI-GAN incorporate saliency or attention mechanisms during training, they still operate on black-box GAN architectures without embedding explainability into the model design. Consequently, the latent space remains entangled and difficult to control at the concept level. This underscores the need for GANs with human-understandable explainability integrated at the architectural level through explicit concept disentanglement and semantically meaningful control and interpretation, especially in domains demanding scientific validity and reliability.

### 3.0 THE CXAI-GAN METHOD

This study proposes CXAI-GAN, a concept-based explainable GAN for plant disease image synthesis that enables independent learning and control of biologically meaningful morphological features.

Table 2: Comparative analysis of standard GANs and the proposed CXAI-GAN for image synthesis.

| Model | Ref. | Generative Learning Method | Disentanglement | Structural Consistency | Explainability | Feature Types |
|---|---|---|---|---|---|---|
| **Original GAN** | [1] | Pixel-level | ✗ Entangled | ✗ Low | ✗ Black-box | ✗ uncontrolled low-level pixels |
| **InfoGAN** | [22] | Pixel-level + mutual information | ✗ Partial | ✗ Weak | ✗ Implicit | ✗ unsupervised latent codes (e.g., width, rotation); lacks concept control or disentanglement |
| **Conditional GAN** | [21] | Pixel + labels | ✗ Partial | ✗ Weak | ✗ Black-box | ✗ Label-driven; coarse class-level control |
| **BigGAN** | [38] | Pixel + labels | ✗ Partial | ✓ Moderate | ✗ Black-box | ✗ Class-conditional, improves visual fidelity; lacks explicit disentanglement and concept control |
| **Pix2Pix** | [23] | Paired pixel-level | ✗ Entangled | ✗ Weak | ✗ Black-box | ✗ Task-specific (e.g., edges to image); no morphological concept |
| **CycleGAN** | [24] | Pixel-level | ✗ Entangled | ✗ Weak | ✗ Black-box | ✗ Domain-to-domain style transfer (e.g., horse, zebra); no concept control |
| **StyleGAN** | [3] | Style-based | ✓ Partial | ✗ Weak | ✗ Limited | ✗ Modifies global appearance traits (e.g., pose, lighting); suffers artifacts |
| **StyleGAN2** | [25] | Style-based | ✓ Partial | ✗ Weak | ✗ Partial | ✗ Pose (position, orientation), lighting, coarse facial features only |
| **StyleGAN3** | [26] | Style-based | ✓ Partial | ✓ Improved | ✗ Partial | ✗ Same features as StyleGAN2, with improved spatial consistency |
| **SC-GAN** | [4] | Style + content | ✓ Partial | ✗ Weak | ✗ Partial | ✗ Content vs. style separation; no domain or biological supervision |
| **GANSpace** | [27] | Style latent directions | ✗ Weak | ✗ Low | ✗ Post-hoc | ✗ Lighting, orientation, smile (via PCA); no concept control |
| **InterFaceGAN** | [28] | Style latent directions | ✗ Weak | ✗ Low | ✗ Post-hoc | ✗ Gender, age, expression (via supervised latent boundaries); |
| **GANterfactual** | [29] | Latent perturbation | ✗ Weak | ✗ Low | ✗ Post-hoc | ✗ Attribute flips via classifier gradients (e.g., smile/no smile); no high-level concept control |
| **xAI-GAN** | [6] | Pixel-level | ✗ Weak | ✗ Low | ✓ Post-hoc | ✗ Saliency and attention maps; no concept disentanglement |
| **ControlNet** | [30] | Diffusion + conditional | ✗ Weak | ✓ High | ✗ Pixel-level control | ✗ Edges, pose, depth maps; strong spatial guidance, no high-level concept control |
| **Diffusion-StyleGAN** | [5] | Diffusion + style latent | ✓ Partial | ✓ High | ✗ Partial | ✗ Generic stylization or class guidance; no high-level concept control |
| **Proposed CXAI-GAN** | | Concept-guided features | ✓ Explicit via CXD generator | ✓ High | ✓ Concept-based explanation | ✓ Independent control over morphological features: shape, texture, and disease patterns |

Concepts refer to high-level attributes such as leaf shape, surface texture, and disease pattern, which reflect how experts interpret and reason about plant disease identification. Unlike low-level pixel or style representations, which are numerical and difficult to interpret, concept-based representations offer explicit control and transparency. CXAI-GAN addresses major limitations of existing GANs, including pixel-level generation, entangled latent spaces, and black-box decision processes that limit their utility in scientific domains.

In contrast to black-box models and style-based approaches like StyleGAN, GANSpace, and InterFaceGAN, CXAI-GAN introduces a Concept-based Explainable and Disentangled (CXD) generator, as shown in Fig. 1, that explicitly encodes and disentangles three high-level morphological features: leaf shape, surface texture, and localized disease patterns. These features are independently disentangled and embedded into a generative pipeline, enabling control over biological attributes. This disentangled design not only facilitates controlled synthesis but also enhances interpretability by aligning the generative process with human-understandable concepts, supporting both scientific transparency and human-in-the-loop integration.

To preserve morphological fidelity and ensure independent control over biological attributes, CXAI-GAN incorporates a fusion embedding mechanism that learns each concept as a distinct representation, including leaf shape, surface texture, and disease pattern. This design enables concept-level interpolation, allowing smooth and interpretable transitions by varying individual conceptual inputs. The generator is explicitly structured to align with high-level biological attributes, shifting from traditional pixel-level generation toward concept-level synthesis. This architecture facilitates the human-centered explainable generation of diverse, high-resolution plant disease images while preserving morphological integrity and biological realism. Each component of the proposed CXAI-GAN is described in detail in the following sections.

### 3.1 Pre-GAN Encoding and Disentangling Latent Subspaces

A core innovation of CXAI-GAN is its ability to disentangle biological attributes prior to the generative process, addressing key limitations in conventional GANs. Instead of using a single entangled latent space, CXAI-GAN encodes the input image into conceptual latent subspaces for leaf shape, texture, and disease patterns. Each subspace corresponds to a semantically meaningful attribute, allowing the generator to learn these factors independently. This structured encoding introduces a semantic representation scheme early in the pipeline, enabling interpretable, controllable, and biologically grounded synthesis. As shown in Equation 1, the encoder $E$ maps the real input image $x_{real}$ into latent subspaces $\{z_s, z_t, z_d\}$ aligned with domain-relevant concepts.

$$E: x_{real} \rightarrow \{z_s, z_t, z_d\} \tag{1}$$

The latent subspace $z_s$ encodes the leaf's morphological and geometric structure, capturing its overall shape, contour variations, and deformations caused by disease progression. The $z_t$ subspace represents surface texture features, including vein patterns, pigment distribution, and other fine-grained details. The $z_d$ subspace encodes disease-specific characteristics such as lesion morphology, discoloration, and spatial spread patterns. By disentangling these attributes into distinct latent subspaces, CXAI-GAN enables precise and independent control over each biological attribute while maintaining semantic clarity and biological fidelity.

#### 3.1.1 Shape Latent Encoding

To generate the shape-disentangled latent subspace, a shape extraction function $O(x)$ creates a binary mask $M_s$ isolating the leaf's structural boundaries while excluding texture and disease details. The masked region $x_s = R(x, M_s)$ captures shape-specific content, which is encoded by $E_s$ into the latent representation $z_s = E_s(x_s)$. As shown in Equation 2 and Fig. 2 (A), this process captures geometric and contour features in a disentangled, shape-specific form.

$$\begin{aligned} M_s &= O(x) \\ x_s &= R(x, M_s) \\ z_s &= E_s(x_s) \end{aligned} \tag{2}$$

#### 3.1.2 Textural Pattern Latent Encoding

To generate the texture-disentangled latent subspace, disease regions are first masked using a disease masking module $M_d$ that removes infected areas based on the disease representation $z_d = E_d(x)$. This produces an intermediate texture image $x'_t$ that isolates the healthy leaf texture. An intensity adjustment operator $M_i$ is then applied to $x'_t$ to simulate pigment and vein intensity variations, enhancing texture diversity. The modified texture $z'_t$ is encoded by $E_t$ into the latent representation $z_t = E_t(z'_t)$, capturing disease-free, disentangled textural features as summarized in Equation 3. The pipeline is illustrated in Fig. 2 (B), where diseased regions are replaced with nearest matching healthy textures.

$$z_d = E_d(x)$$
$$x'_t = M_d(x, z_d)\ ,\ z'_t = M_i(x'_t) \tag{3}$$
$$z_t = E_t(z'_t)$$

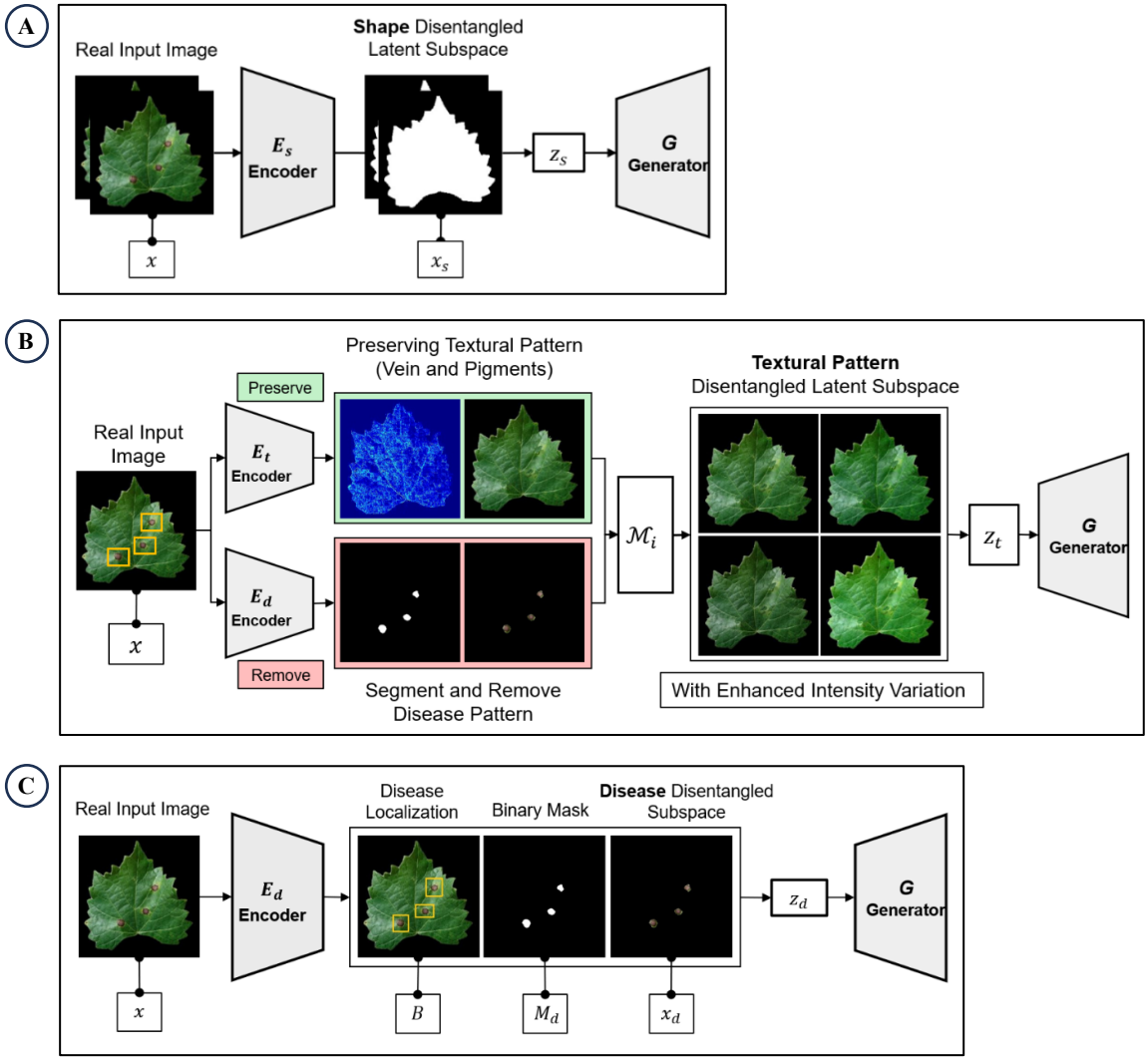


Fig. 2: Encoding pipeline for disentangling latent subspaces: (A) Shape-specific encoding, (B) Texture pattern encoding, and (C) Disease-specific encoding.

### 3.1.3 Disease Pattern Latent Encoding

To generate the disease-disentangled latent subspace, disease-affected regions in the input image are localized using a detection function $D(x)$, which creates bounding boxes $B$. These regions are segmented to form a disease mask $M_d$, isolating the pathological content $x_d = R(x, M_d)$. An encoder $E_d$ then maps this isolated region into the latent representation $z_d$, capturing disease-specific features independently from other attributes. This encoding process is formalized in Equation 4 and illustrated in Fig. 2 (C), showing the localization, segmentation, and latent mapping of disease patterns.

$$B = D(x)\ ,M_d = S(x, B)$$
$$x_d = R(x, M_d) \tag{4}$$
$$z_d = E_d(x_d)$$

### 3.2 CXAI-Guided Generator Training for Interpretable Image Synthesis

The Concept-based Explainable and Disentangled (CXD) generator forms the core of the proposed CXAI-GAN architecture. It distinguishes itself from conventional GANs by explicitly incorporating disentangled and semantically grounded feature learning into the generation process, as illustrated in Fig. 3. Unlike standard GANs that rely on dense, entangled latent vectors where shape, texture, and disease features are mixed in an unstructured fashion, CXAI-GAN introduces a disentangled latent design that separates key biological attributes into three interpretable subspaces: leaf shape, textural pigmentation, and disease-localized patterns.

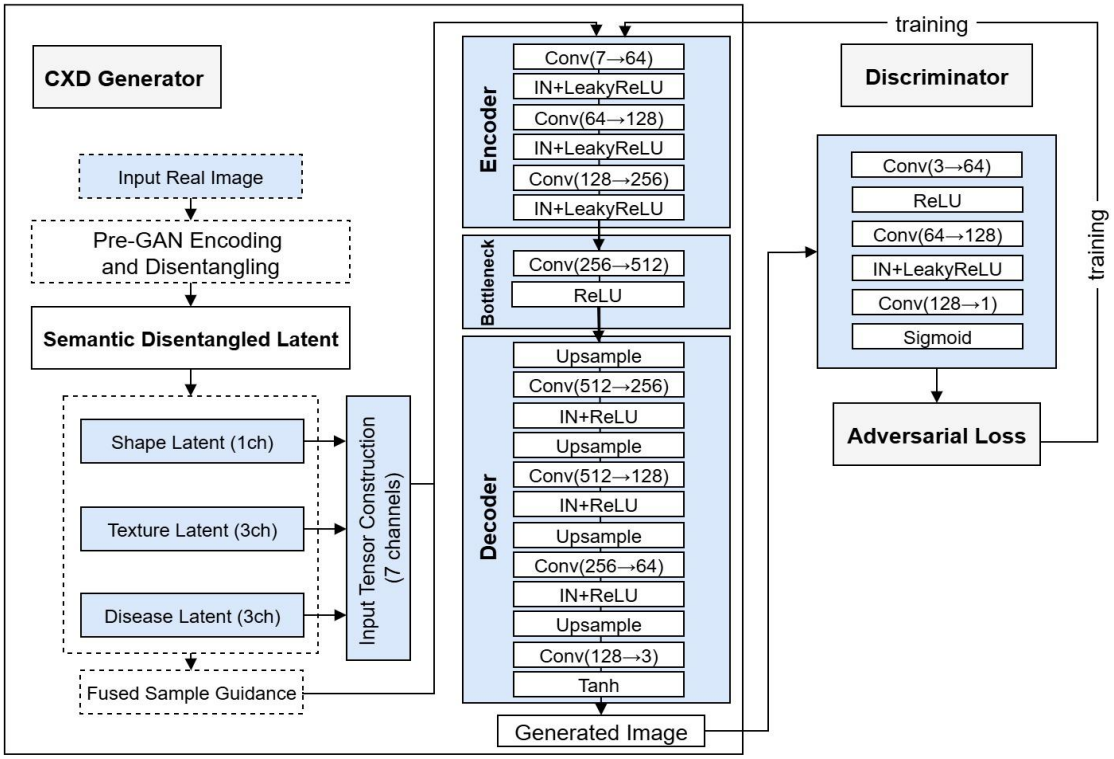


Fig. 3: Block diagram of the proposed CXAI-GAN model, illustrating the Concept-based Explainable and Disentangled (CXD) Generator. The generator receives disentangled representations of shape, texture, and disease attributes independently to synthesize biologically realistic and semantically interpretable images.

The CXD Generator adopts a U-Net-based architecture with skip connections to maintain spatial integrity and local detail throughout the generation pipeline. These architectural elements ensure that fine-grained structural features such as venation and lesion boundaries are preserved while enabling attribute-specific control. The generator receives three independently encoded latent vectors corresponding to the disentangled subspaces and maps them to high-resolution synthetic leaf images through a structured decoding process. This process is mathematically represented as:

$$G(z_s, z_t, z_d) = F(F_s(z_s), F_t(z_t), F_d(z_d)) \tag{5}$$

where $z_s$, $z_t$, and $z_d$ represent latent codes for shape, texture, and disease, respectively; $F_s$, $F_t$, and $F_d$ denote their corresponding decoders; and $F$ denotes the fusion mechanism that combines the disentangled features into a unified visual representation. This disentangled mapping supports controlled synthesis, allowing explicit manipulation of biological features, capabilities not feasible in conventional black-box, pixel-level GANs.

Training is conducted via an adversarial learning loop involving a convolutional discriminator that distinguishes real from generated images and enforces statistical alignment with real-world distributions. This adversarial feedback enhances visual realism, while the disentangled architecture ensures interpretability and biological plausibility. In contrast to conventional GANs, which often suffer from entangled representations, mode collapse, and weak semantic control, CXAI-GAN enables fine-grained, interpretable generation of disease-relevant features and advances its utility in biologically grounded tasks such as phenotyping, classification, and in silico experimentation.

This study utilizes the PlantVillage dataset [39], a widely recognized benchmark for plant disease diagnosis tasks. A training set of 2,000 images was used to train the model over 200 epochs with a batch size of 8 on 128×128

resolution images, using the Adam optimizer with a learning rate of 0.0002. To generate synthetic data, we first selected a subset of 100 real grape Black Rot leaf samples and encoded their conceptual representations in a disentangled latent space. Specifically, 100 texture embeddings were extracted and expanded with intensity variations to form 400 distinct texture representations, along with 100 disease pattern embeddings. These were interpolated with only five distinct shape embeddings, which were held constant to manage dataset scalability and avoid unnecessary redundancy (2,000 ← 400*5). Although more shape variants could be used, we empirically selected five to ensure sufficient diversity while maintaining a manageable dataset size. These disentangled features were then recombined via structured embedding. For instance, a fixed leaf shape could be paired with varying texture and disease patterns, allowing controlled manipulation of individual biological traits essential for interpretability and effective training.

To ensure meaningful supervision during training, we constructed paired data samples for each disentangled component by systematically combining them into structured triplets. These paired combinations allowed the generator to learn fine-grained associations between latent codes and their visual outputs. In addition to these individual combinations, we also created fused samples by integrating all three components prior to generative learning, providing guidance to the generator. This approach enhanced the generator's ability to synthesize biologically coherent images while preserving control over individual traits. It strengthened the model's capacity to generate biologically faithful images and ensured fine-grained control over each independent feature.

### 3.3 CXAI-GAN Inference

During inference, CXAI-GAN can generate new synthesized images through interpolation within disentangled latent subspaces corresponding to biologically meaningful attributes: leaf shape $(z_s)$, texture $(z_t)$, and disease patterns $(z_d)$. This interpolation of conceptual features enables independent, semantically consistent modulation of each attribute, producing morphologically and biologically plausible variations. The explicit disentanglement preserves critical features, supporting fine-grained control essential for modelling disease progression. Furthermore, the concept-aligned latent representations enable human-in-the-loop interaction, allowing domain experts to iteratively refine latent vectors based on expert knowledge. This interactive process enhances interpretability, adaptability, and contextual alignment of synthetic outputs, directly supporting biologically informed tasks such as data augmentation and expert-guided phenotyping.

Fig. 4 illustrates the concept-level embedding interpolation strategy using the fusion method, highlighting the transition from independently encoded conceptual features to the explainable synthesis of output images. This figure demonstrates the capability of interpolating each disentangled feature independently; for instance, the shape latent representation can be held constant while varying texture patterns from different leaves, enabling controlled and biologically meaningful synthesis of diverse image variations.

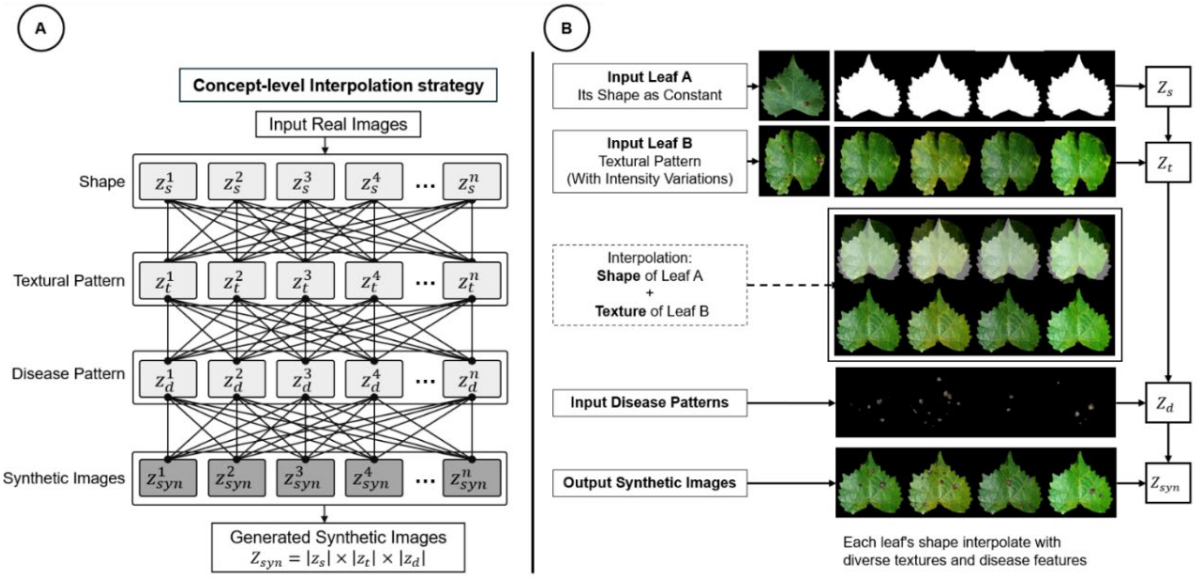


Fig. 4: A) Concept-based embedding interpolation strategy. B) Illustration of independent interpolation within disentangled subspaces for controlled and explainable image synthesis.

## 4.0 EVALUATION

To comprehensively evaluate the proposed CXAI-GAN's performance to generate high-quality synthetic images and assess its practical utility in downstream machine learning tasks, we conduct four distinct evaluations based on standard GAN benchmarks and CNN-based metrics. These include Quantitative Quality Evaluation, Fine-Grained Feature Evaluation, Spatial Consistency Evaluation, and Downstream Task Performance.

### 4.1 Quantitative Quality Evaluation

To evaluate the quality of synthetic images generated by proposed model, we employ three widely-used metrics: Fréchet Inception Distance (FID) [40], Structural Similarity Index (SSIM) [41], and Peak Signal-to-Noise Ratio (PSNR) [42]. FID measures the distributional similarity between real and generated images in a feature space, effectively capturing perceptual quality and diversity. SSIM assesses the structural similarity by comparing luminance, contrast, and texture between images, providing a comprehensive measure of visual fidelity. PSNR quantifies the pixel-level reconstruction quality by evaluating the ratio between the maximum possible signal and the noise affecting the image, offering insight into overall image fidelity. Together, these metrics provide a robust and complementary evaluation of the realism and quality of the generated images.

$$\begin{aligned} FID &= \| \mu_r - \mu_g \|_2^2 + Tr\left(\sum_r + \sum_g - 2(\sum_r \sum_g)^{1/2}\right) \\ SSIM(x,y) &= \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \\ PSNR &= 10 \,.\, log_{10}\left(\frac{MAX_I^2}{MSE}\right) \end{aligned} \tag{6}$$

In FID, $\mu_r$, $\sum_r$ and $\mu_g$, $\sum_g$ are the mean and covariance of real and generated image features, respectively. In SSIM, $\mu_x$, $\mu_y$ are means, $\sigma_x^2$, $\sigma_y^2$ are variances, and $\sigma_{xy}$ is the covariance of images $x$ and $y$. In PSNR, $MAX_I$ is the maximum possible pixel value of the image and $MSE$ is the mean squared error between the images.

### 4.2 Fine-Grained Feature Evaluation

To assess textural and structural fidelity, we utilize the Vein-Pigment Similarity (VPS) metric, which evaluates the correspondence between vein patterns and pigment distributions, key biological features for plant disease characterization. VPS is calculated by extracting histograms of vein and pigment intensities from real and synthetic images and computing the correlation between these histograms. Specifically, VPS measures the similarity as the correlation between histogram values $h_1(i)$ and $h_2(i)$ at each bin $i$, thereby quantifying how closely the synthetic images replicate the natural alignment of pigment intensity gradients and venation structures. Elevated VPS scores indicate superior preservation of biologically relevant texture features, critical for accurate disease modeling.

$$VPS = \frac{\sum_i \quad h_1(i).\, h_2(i)}{\sqrt{\sum_i \quad h_1(i)^2 \;.\; \sum_i \quad h_2(i)^2}} \tag{7}$$

### 4.3 Spatial Consistency Evaluation

To evaluate the spatial alignment of structural features, we use the Hausdorff Distance Similarity (HDS) [43], which measures geometric correspondence between binary regions, specifically the leaf outline contours in real and synthetic images. As defined in Equation 8, HDS computes the maximum of the minimum Euclidean distances between point sets $A$ and $B$, which represent the boundary points of real and synthetic leaf shapes, respectively. Here, $d(a,b)$ and $d(b,a)$ denote the shortest distances from a point in one set to its nearest neighbor in the other. The metric is normalized so that higher scores indicate better spatial alignment, reflecting how accurately synthetic images preserve real anatomical structures, crucial for reliable plant disease analysis.

$$\begin{aligned} H(A,B) &= max\,(h(A,B), h(B,A) \\ h(A,B) &= min\,(d(a,b)) \mid a \in A \text{ and b} \in B \\ h(B,A) &= min\,(d(b,a)) \mid \text{b} \in B \text{ and } a \in A \end{aligned} \tag{8}$$

### 4.4 Downstream Task Performance

To validate the practical utility of CXAI-GAN, we evaluate its impact on a downstream leaf disease classification task using three training scenarios: (1) real-only data, (2) real data augmented with generated synthetic images, and (3) synthetic-only data. A classification model is trained under each setting and evaluated on a real test set to assess generalization. Performance is measured using classification accuracy, Precision, and F1-score. This comparative analysis highlights the effectiveness of CXAI-GAN in generating biologically meaningful samples that can either complement real data or, in some cases, serve as a stand-alone resource for model training.

## 5.0 RESULTS

### 5.1 Quantitative Quality Results

To quantitatively assess the realism and fidelity of images generated by CXAI-GAN, we report three standard metrics: FID, SSIM, and PSNR. As shown in Table 3, CXAI-GAN achieves a FID of 19.64, indicating strong alignment between the synthetic and real image distributions in the deep feature space. This reflects the model's ability to generate biologically realistic images that are statistically consistent with natural data. The SSIM score of 0.955 reflects excellent structural similarity, effectively preserving morphological details such as leaf venation and disease lesion boundaries, key factors for semantic fidelity in biological contexts. Additionally, the PSNR of 34.91 dB confirms high visual fidelity and minimal pixel-level distortion, demonstrating that the model accurately reconstructs fine-grained content. These results collectively validate the effectiveness of CXAI-GAN's concept-guided disentanglement and synthesis framework in producing high-quality, biologically meaningful plant disease images.

Table 3: Quantitative evaluation of the generated images using SSIM, FID, and PSNR on the best performance. The results indicate high structural similarity, biological realism, and visual fidelity of the proposed CXAI-GAN.

| **Metric** | **Score** | **Interpretation** |
|---|---|---|
| **SSIM** | 0.955 | High structural similarity of morphological and biological details. |
| **FID** | 19.64 | Good ability to generate biologically realistic images. |
| **PSNR** | 34.91 | High visual fidelity and fine-grained accuracy. |

To assess potential overfitting, we monitored the training progression of both the generator (G) and discriminator (D) losses over 200 epochs. As illustrated in Fig. 5, the losses show stable and consistent improvement without sudden spikes or divergence. Concurrently, image quality metrics such as SSIM and PSNR steadily increase, while FID scores remain within an acceptable range, indicating close alignment between generated and real data distributions. These stable trends confirm that CXAI-GAN generalizes well, capturing biologically relevant features without overfitting or memorizing the training samples.


Fig. 5: Progression of quantitative metrics (SSIM, FID, PSNR) across training epochs. The results demonstrate consistent improvements in structural similarity, visual fidelity, and biological realism as training advances.

### 5.2 Fine-Grained and Spatial Results

This section presents the evaluation of specific biological features, focusing on fine-grained texture and spatial structure, using domain-aware metrics: Vein-Pigment Similarity (VPS), Hausdorff Distance Similarity (HDS), and local SSIM. In addition to general image quality metrics such as FID, SSIM, and PSNR that assess global perceptual similarity, these metrics individually measure the fidelity of disentangled and interpretable features. The quantitative evaluation results achieved a mean VPS score of 0.999, a mean HDS of 0.973, and a mean local SSIM of 0.937, indicating strong preservation of venation, pigment textures, and spatial leaf geometry, respectively.

These results, as shown in Table 4, confirm the model's ability to generate biologically plausible features with fine-grained control, underscoring the effectiveness of the concept-guided disentangled architecture. As illustrated in Fig. 6, the distribution histograms of SSIM, VPS, and HDS scores further support this claim by showing consistently high values across all samples. Additionally, Fig. 7 provides a visual comparison between real and synthetic features, highlighting the model's ability to generate interpretable and biologically realistic outputs in terms of both texture and shape fidelity.

Table 4: Quantitative evaluation of individual disentangled biological features in synthetic images.

| **Metric** | **Mean** | **Std. Dev.** | **Interpretation** |
|---|---|---|---|
| **HDS** | 0.973 | 0.020 | Strong shape similarity, indicating accurate disentangled geometry. |
| **VPS** | 0.999 | 0.0009 | High texture alignment, validating disentangled vein and pigment synthesis. |
| **Local SSIM** | 0.937 | 0.031 | High local structural similarity, confirming fine feature preservation. |
| **Mean (All)** | 0.970 | 0.014 | Overall high fidelity across disentangled biological features. |

### 5.3 Downstream Classification Results

To evaluate the practical effectiveness of images generated by CXAI-GAN, we trained a CNN classifier to distinguish between two visually similar grape diseases: Esca and Black-Rot. As shown in Table 5, training on real data achieved a test accuracy of 96.7% for Esca but only 86.7% for Black-Rot, highlighting the challenge of capturing subtle inter-class variations.

The comparable performance between the real-only and real + synthetic setups demonstrates that the synthetic data generated by CXAI-GAN is highly aligned with the real data distribution. Rather than introducing noise or domain shift, the synthetic images reinforce the discriminative features already present in the real data. This

consistency supports the conclusion that CXAI-GAN produces biologically faithful, high-quality samples suitable for downstream tasks.

Notably, the model trained on synthetic data achieves performance comparable to the real-only setup, with 96.7% accuracy for both Esca and Black-Rot. This reflects a 10% improvement in Black-Rot classification accuracy over the other training setups, strongly validating the utility of the generated synthetic images. These results demonstrate that CXAI-GAN can effectively supplement or even substitute real data in scenarios with limited annotations, such as rare disease phenotypes. As illustrated in Fig. 8, these are examples of images that were misclassified by the classifier trained on real data, but correctly classified when using a classifier trained and augmented with synthetic images generated by the proposed CXAI-GAN.

Table 5: Classification performance under different training configurations. Augmenting real data with CXAI-GAN's generated synthetic samples improved Black-Rot accuracy from 86.7% to 96.7% with 10% improvement, demonstrating enhanced feature diversity and generalization across visually similar and challenging disease classes.

| | **Training** | | | | **Validation** | | | | **Test** | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Training Setup** | **Loss** | **Accuracy** | **F1-Score** | **Precision** | **Loss** | **Accuracy** | **F1-Score** | **Precision** | **Accuracy** (Esca) | **Accuracy** (Black-Rot) |
| **Real-Only** | 0.18 | 0.93 | 0.93 | 0.93 | 0.07 | 0.98 | 0.99 | 0.98 | 96.7 | **86.7** |
| **Real + Synthetic** | 0.17 | 0.93 | 0.93 | 0.93 | 0.06 | 0.99 | 0.99 | 0.99 | 96.7 | **93.3** |
| **Synthetic-Only** | 0.11 | 0.96 | 0.96 | 0.96 | 0.03 | 0.99 | 0.99 | 0.99 | 96.7 | **96.7** |

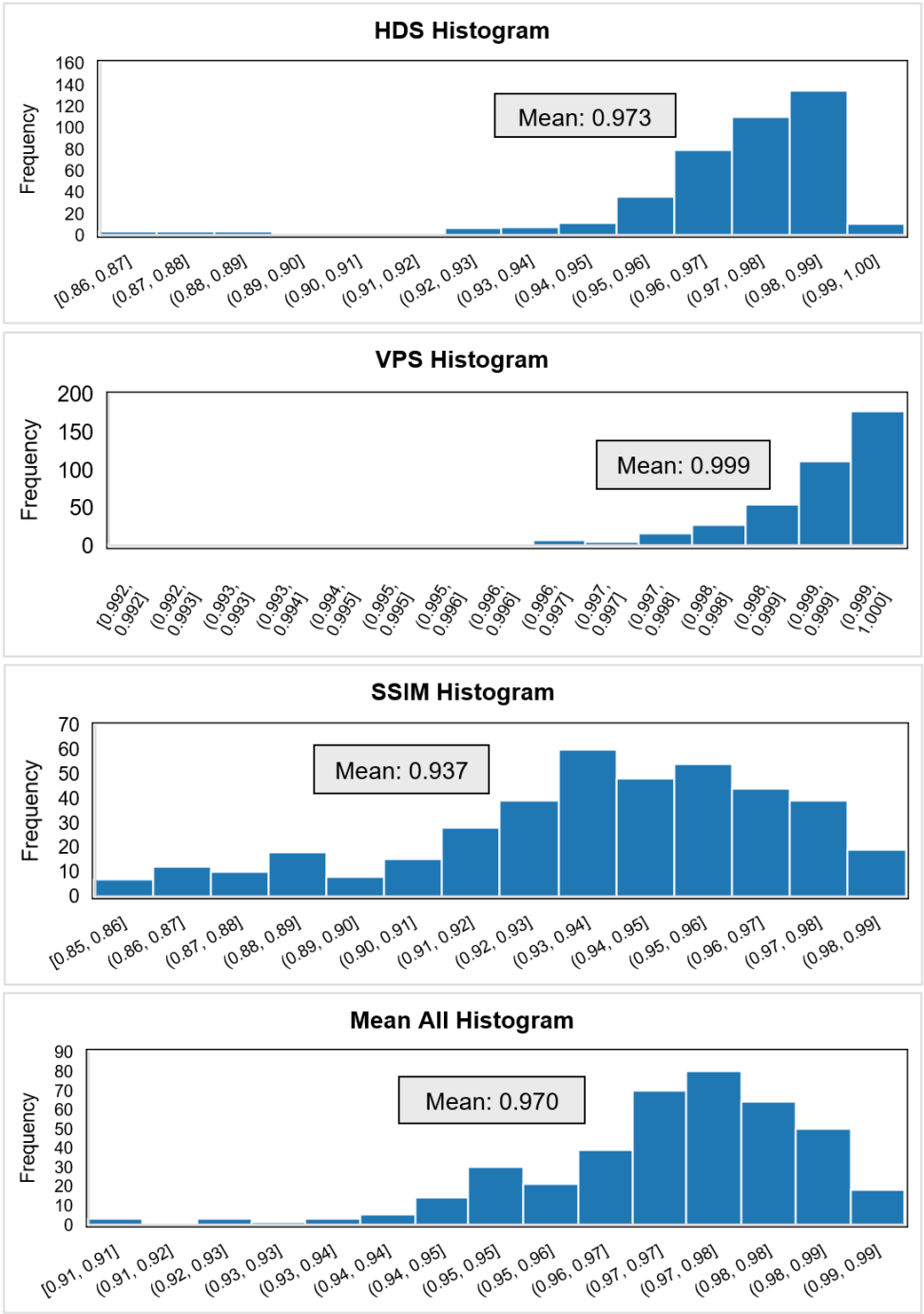


Fig. 6: Distribution histograms of SSIM, VPS, and HDS scores. Higher values across these metrics indicate strong structural similarity (SSIM), accurate reproduction of texture patterns (VPS), and spatial alignment of leaf shapes (HDS), validating the fidelity of disentangled biological features in the generated images.

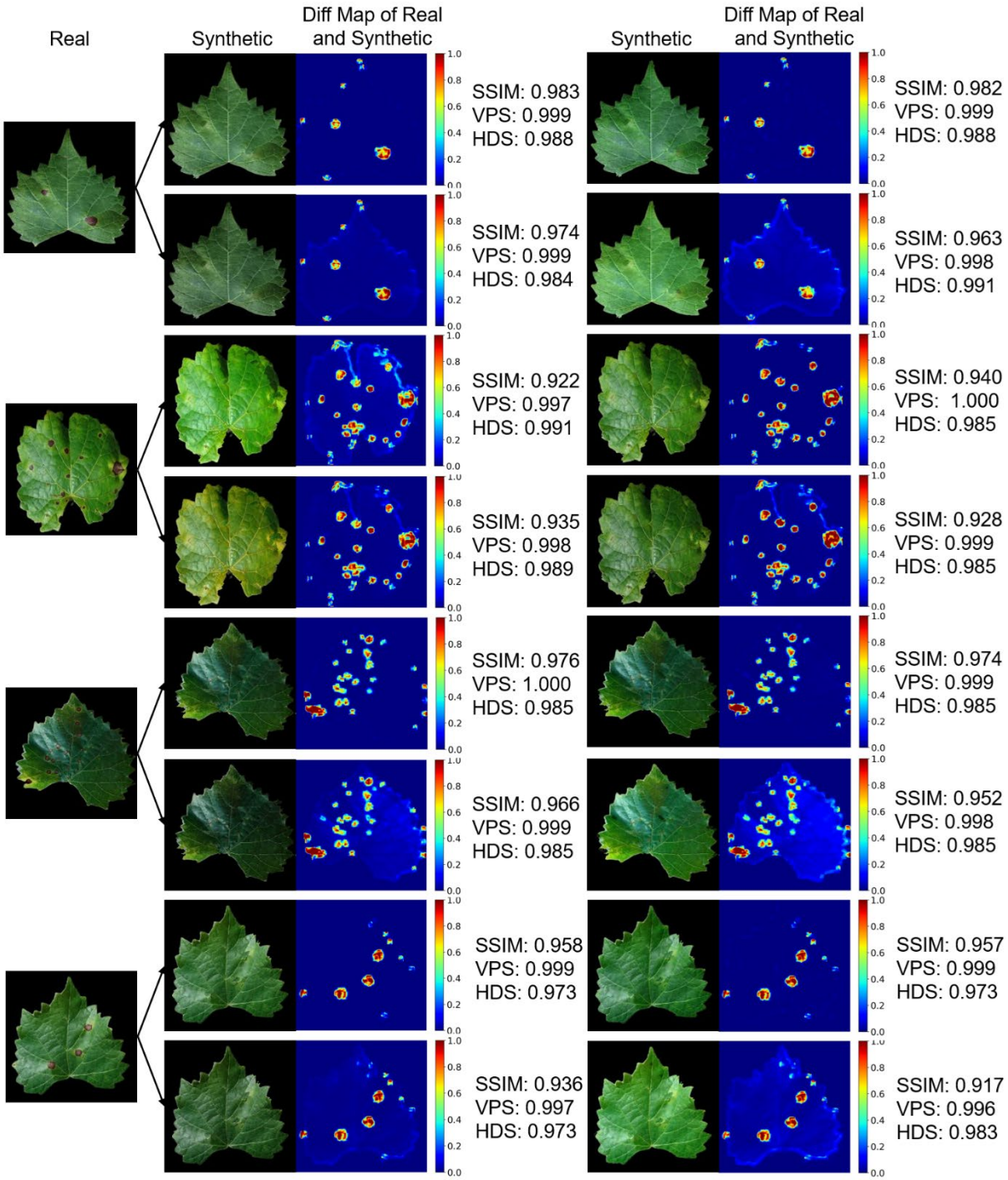


Fig. 7: Comparison of real and synthetic disentangled features. Evaluated using VPS, HDS, and SSIM metrics; high scores confirm the model's ability to generate biologically meaningful and interpretable features, including fine-grained textures and spatial leaf structures.

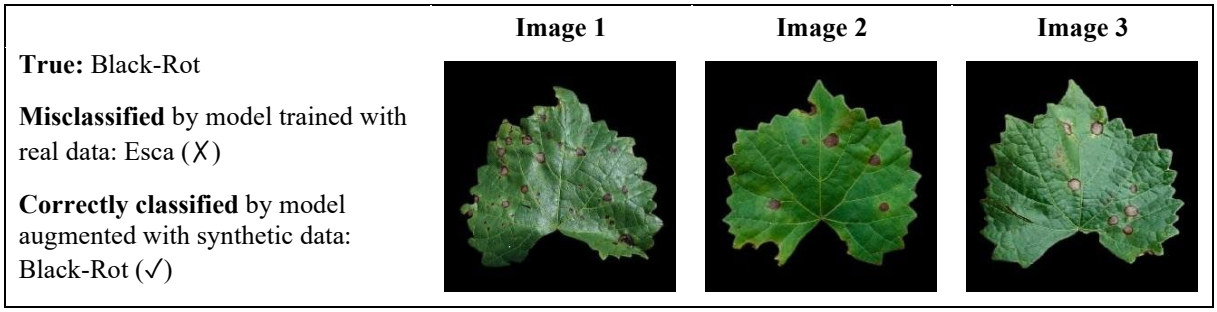


Fig. 8: Examples of images misclassified by the classifier trained on real data only but correctly classified by the classifier trained on synthetic images from proposed CXAI-GAN.

## 6.0 DISCUSSION

### 6.1 Advancing Explainable Generative Modelling

This study introduces CXAI-GAN, a concept-based and explainable generative adversarial network tailored for biologically realistic plant disease image synthesis. The core challenge addressed by CXAI-GAN lies in the inherent limitations of standard GANs, stemming from pixel-level generation and black-box architectures. Early GANs operate primarily at the pixel level, capturing low-level correlations without encoding semantic structure, which leads to entangled latent representations that constrain interpretability and limit meaningful control over synthesized outputs.

InfoGAN introduced an unsupervised method to promote disentanglement by maximizing mutual information between latent codes and generated outputs, enabling the model to learn limited interpretable factors such as rotation and width without supervision. However, this disentanglement is implicit, coarse, and often failing to capture semantically meaningful or biologically relevant features in specialized fields such as plant disease identification. In contrast, CXAI-GAN explicitly disentangles and controls high-level biological concepts, shape ($z_s$), texture ($z_t$), and disease characteristics ($z_d$) through supervised, domain-informed learning via the Concept-based Explainable and Disentangled (CXD) Generator. This structured concept-level representation enhances controllability and supports explainable AI and human-in-the-loop interpretability, allowing experts to understand and guide generation based on meaningful attributes. Unlike InfoGAN's latent-only interpretability, CXAI-GAN offers a transparent, modular, and biologically aligned framework for fine-grained, semantically rich image synthesis.

While StyleGAN enables moderate semantic modulation via intermediate latent injection, it lacks explicit feature supervision, resulting in a semantically entangled and opaque latent space that restricts its suitability for interpretable and structured synthesis in biological domains. Recent StyleGAN-based methods, such as GANSpace and InterFaceGAN, offer post-hoc interpretability by uncovering latent directions that correspond to limited semantic variations within pretrained StyleGANs. These approaches enable modification of coarse visual attributes such as lighting or expression. For example, GANSpace applies unsupervised PCA to discover directions that adjust pose or illumination, while InterFaceGAN uses supervised classifiers to define decision boundaries for features like age or gender. However, these methods are limited by their reliance on entangled latent representations and lack explicit conceptual representation learning of morphological and pathological features such as infected leaf geometry, venation, and pigment distribution, or disease region structure.

Hybrid models that combine StyleGAN with diffusion priors enhance high-fidelity image quality but still lack semantically grounded disentanglement. In contrast, xAI-GAN, which does not rely on StyleGAN, incorporates attention mechanisms and saliency-based methods such as Saliency Maps, LIME, and SHAP to improve interpretability during training. However, xAI-GAN treats explainability primarily as a post-hoc visualization tool, rather than embedding explicit concept-level learning into the generative process. This limits its effectiveness as a human-understandable model and restricts its capability for human-in-the-loop applications in scientific image synthesis.

These limitations are particularly restrictive for explainable GAN models in biological image synthesis, where pixel-level accuracy alone is insufficient. Applications such as plant disease modelling require generative systems with explicit, modular control over morphological and pathological features to support tasks like data augmentation, disease progression simulation, and expert-in-the-loop analysis. CXAI-GAN addresses this need by incorporating concept-level disentanglement and semantic supervision into the generation process, enabling both transparent interpretation and biologically faithful synthesis.

CXAI-GAN establishes a benchmark in explainable generative modelling for biological image synthesis by combining high-fidelity generation with concept-level disentanglement. It learns independently controllable latent codes for shape, texture, and disease pattern through concept-level supervision, enabling precise manipulation and interpolation of each biological attribute. This design supports the generation of novel yet biologically faithful images, maintaining both structural integrity and interpretability.

Quantitative evaluations affirm the superiority of CXAI-GAN across multiple dimensions. The model achieves an FID of 19.64, reflecting a close alignment with the real data distribution, alongside an SSIM of 0.955 and PSNR of 34.91 dB, indicating strong perceptual and pixel-level fidelity. For fine-grained biological assessment, it reaches a Hausdorff Distance Similarity (HDS) of 0.973, a Vein-Pigment Similarity (VPS) of 0.999, and local SSIM scores of 0.937 across shape and texture, confirming accurate preservation of localized biological traits. To assess practical utility in downstream tasks, we evaluated the synthesized images in three classification settings: (i) real-only training, (ii) hybrid training with real and synthetic images, and (iii) synthetic-only training. In the synthetic-only scenario, CXAI-GAN generated images achieved an F1-score of 0.96, precision of 0.96, and precision of 0.96, demonstrating that its generated images are not only visually convincing but also semantically informative. Notably, the hybrid training setup improved generalization and mitigated overfitting, suggesting CXAI-GAN's ability to complement real datasets.

Beyond quantitative performance, CXAI-GAN's transparent generation pipeline supports critical applications such as disease detection, data augmentation, and human-in-the-loop validation, where interpretability and controllability are essential. By embedding domain-informed conceptual constraints into the latent space learning process, CXAI-GAN aligns with the emerging paradigm of concept-based explainable AI (XAI), which emphasizes human-understandable abstractions to facilitate collaborative scientific reasoning and human-AI workflows. This is especially vital in biomedical and regulatory contexts, where transparency and auditability are prerequisites for expert trust, safety validation, and compliance with standards such as those set by the FDA. In this context, CXAI-GAN goes beyond visual realism to enable interpretable, trustworthy, and practically useful generative modelling for real-world biological applications.

### 6.2 Human-in-the-Loop Capability

While CXAI-GAN is primarily developed as an explainable generative model, its architecture inherently supports human-in-the-loop (HIL) workflows, enabling effective human-AI interaction. The model's disentangled representation learning for biologically meaningful concepts, such as leaf shape, texture, and disease pattern, provides interpretable controls that experts can manipulate to generate targeted synthetic images. This facilitates iterative refinement, validation, and expert feedback loops critical for scientific discovery and applied agricultural research.

Incorporating real-time expert feedback can improve model robustness and promote trust by making the AI system responsive to domain knowledge. Such human-AI collaboration can leverage active learning strategies where the model prioritizes uncertain or challenging cases for expert input, thereby enhancing learning efficiency and reducing bias. Additionally, embedding ethical AI principles into these interactions is essential to ensure fairness, transparency, and accountability, especially in high-stakes biological applications where decisions may impact livelihoods and food security. Although this study does not implement a full HIL system, CXAI-GAN's modular design lays the groundwork for future integration of interactive feedback, active learning, and ethical AI.

### 6.3 Limitations and Future Work

CXAI-GAN presents a concept-driven generative framework that combines semantic interpretability with high-fidelity image synthesis in the context of plant disease modelling. While the current design effectively disentangles three core biological features: leaf shape, surface texture, and disease pattern, its conceptual scope remains limited to a predefined set of attributes. A natural direction for future work involves expanding this semantic space to incorporate additional morphological and pathological descriptors. For example, attributes such as lesion size and severity, symptom spatial distribution (such as clustered or scattered patterns), vein structure, or pigment variation could provide more detailed explanation and enhance relevance for researchers and practitioners.

Moreover, the current framework is trained and evaluated on a 2D infected leaf image dataset. While this setup supports rigorous validation and interpretability, it may not fully represent the diversity of real-world agricultural environments. Extending the model to handle more complex data types such as multi-angle imaging, temporal sequences of disease development, or complementary modalities like hyperspectral imagery would improve its robustness and applicability in practical settings. These future directions remain aligned with the model's core emphasis on explainability, allowing it to support more comprehensive simulations and downstream diagnostic tasks without compromising interpretability.

In addition to expanding the model's conceptual and data scope, integrating human-in-the-loop (HIL) capabilities represents a critical avenue for future development. Enabling real-time expert interaction with CXAI-GAN would allow users to guide the generative process, provide corrective feedback, and iteratively refine outputs based on domain knowledge. Such collaboration not only enhances model accuracy and relevance but also fosters

trust and transparency essential for deployment in scientific and regulatory environments. Incorporating active learning strategies within this interactive framework could further optimize training efficiency by prioritizing expert-reviewed samples. Emphasizing ethical AI principles throughout this process will ensure that the system remains accountable, fair, and aligned with human values as it transitions toward practical, real-world applications. Collectively, these advancements will evolve CXAI-GAN from a specialized prototype into a robust, biologically grounded generative platform. This will expand its utility across diverse applications such as plant phenotyping, disease monitoring, and decision-support systems, domains where transparency and interpretability are critical for adoption by researchers, practitioners, and regulatory bodies.

## 7.0 CONCLUSION

Biological image synthesis presents distinct challenges, including the need for morphological precision, expert trust, and transparency. Although existing GANs can produce visually high-fidelity images, they typically rely on pixel-level generation and entangled latent representations, which offer limited interpretability. These drawbacks lead to a lack of morphological control, limiting the applicability of existing GANs in scientific and clinical domains where understanding a model's internal mechanisms and achieving human-understandable image synthesis are critical.

In this study, we propose CXAI-GAN, a concept-based explainable generative adversarial network that disentangles biologically morphological features, such as infected leaf shape, surface texture, and disease pattern, into independent latent subspaces. Unlike existing GAN models like StyleGAN, GANSpace, and hybrid diffusion StyleGAN, CXAI-GAN provides explicit morphological control through a disentangled generator guided by high-level concepts. This design facilitates robust morphological representation learning, enhances biological relevance, and builds the generative process more transparent and interpretable.

This work aligns with a broader evolution in explainable AI (XAI), moving from post-hoc interpretability tools toward inherently interpretable architectures. In particular, concept-based explainable AI (C-XAI) has emerged as a promising direction that incorporates human-centered concepts. By embedding these concepts directly into the learning and generation process, C-XAI enables a closer alignment between black-box machine learning model functionality and expert understanding, which is especially critical in high-stakes domains such as plant disease detection, and plant health monitoring. CXAI-GAN demonstrates the practical value of this paradigm. Its human-interpretable generative process supports human-in-the-loop analysis and enables collaborative human-AI workflows, where experts can guide, inspect, and validate outputs in real time. This concept-guided generation approach enables explainable image synthesis and contributes to the development of trustworthy, expert-centered AI systems in agriculture and scientific domains.

The model achieved strong quantitative performance, with an FID of 19.64, SSIM of 0.955, and PSNR of 34.91. In a classification task involving two visually similar grape diseases, augmentation with CXAI-GAN images improved test accuracy by 10%, and synthetic-only training achieved 96.7% accuracy. These results confirm that the model produces high-quality, biologically meaningful images suitable for downstream analysis and expert validation. Current limitations include reliance on 2D images and a limited set of morphological concepts. Future work could focus on extending CXAI-GAN to support time-series data, incorporating a broader range of biological features, and integrating adaptive expert feedback for iterative model refinement.

## REFERENCES


[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *28th International Conference on Neural Information Processing Systems,* vol. 27, pp. 2672-2680, 2014.

[2] Z. ur Rahman, M. S. M. Asaari, H. Ibrahim, I. S. Z. Abidin, and M. K. Ishak, "Generative Adversarial Networks (GANs) for Image Augmentation in Farming: A Review," *IEEE Access,* pp. 179912-179943, 2024, doi: 10.1109/ACCESS.2024.3505989.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401-4410.

[4] H. Kazemi, S. M. Iranmanesh, and N. Nasrabadi, "Style and content disentanglement in generative adversarial networks," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019: IEEE, pp. 848-856.

[5] H. Zhang, F. Yao, Y. Gong, and Q. Zhang, "Anemone Image Generation Based on Diffusion-Stylegan2," *IEEE Access,* vol. 12, pp. 37310-37325, 2024.

[6] V. Nagisetty, L. Graves, J. Scott, and V. Ganesh, "xai-gan: Enhancing generative adversarial networks via explainable ai systems," *arXiv preprint arXiv:2002.10438,* 2020.

[7] X. Qin, F. M. Bui, Z. Han, and A. Khademi, "Toward improved interpretability in medical imaging: Revealing the disease evidence from chest X-Ray images using an adversarial generative approach," *IEEE Access,* vol. 12, pp. 82002-82014, 2024.

[8] D. Saxena and J. Cao, "Generative adversarial networks (GANs) challenges, solutions, and future directions," *ACM Computing Surveys (CSUR),* vol. 54, no. 3, pp. 1-42, 2021.
[9] X. Zheng, X. Yang, Q. Zhao, H. Zhang, X. He, J. Zhang, and X. Zhang, "CFA-GAN: Cross fusion attention and frequency loss for image style transfer," *Displays,* vol. 81, p. 102588, 2024.
[10] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "Gans for medical image synthesis: An empirical study," *Journal of Imaging,* vol. 9, no. 3, p. 69, 2023.
[11] A. A. Showrov, M. T. Aziz, H. R. Nabil, J. R. Jim, M. M. Kabir, M. Mridha, N. Asai, and J. Shin, "Generative adversarial networks (GANs) in medical imaging: advancements, applications and challenges," *IEEE Access,* 2024.
[12] S. A. Atone and A. Bhalchandra, "Generative adversarial networks in computer vision: a review of variants, applications, advantages, and limitations," in *International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering & Technology*, 2023: Springer, pp. 589-608.
[13] Z. Ahmad, Z. u. A. Jaffri, M. Chen, and S. Bao, "Understanding GANs: Fundamentals, variants, training challenges, applications, and open problems," *Multimedia Tools and Applications,* pp. 1-77, 2024.
[14] Q. Li, Y. Tang, and L. Chu, "Generative adversarial networks for prognostic and health management of industrial systems: A review," *Expert Systems with Applications,* p. 124341, 2024.
[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "" Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135-1144.
[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *31st International Conference on Neural Information Processing Systems,* vol. 30, pp. 4768-4777, 2017.
[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization. arXiv: 161002391v4," *Explan. Vis. Networks, Deep Localization, Gradient-based Networks, Deep,* pp. 1-22, 2019.
[18] T. N. Mundhenk, B. Y. Chen, and G. Friedland, "Efficient saliency maps for explainable AI," *arXiv preprint arXiv:1911.11293,* 2019.
[19] S. Natarajan, S. Mathur, S. Sidheekh, W. Stammer, and K. Kersting, "Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, no. 27, pp. 28594-28600.
[20] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: a state of the art," *Artificial Intelligence Review,* vol. 56, no. 4, pp. 3005-3054, 2023.
[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784,* 2014.
[22] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *30th International Conference on Neural Information Processing Systems,* vol. 29, pp. 2180-2188, 2016.
[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125-1134.
[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223-2232.
[25] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110-8119.
[26] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in neural information processing systems,* vol. 34, pp. 852-863, 2021.
[27] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," *Advances in neural information processing systems,* vol. 33, pp. 9841-9850, 2020.
[28] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9243-9252.
[29] S. Mertes, T. Huber, K. Weitz, A. Heimerl, and E. André, "Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning," *Frontiers in artificial intelligence,* vol. 5, p. 825565, 2022.
[30] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836-3847.
[31] Z. Shams Khoozani, A. Q. M. Sabri, W. C. Seng, M. Seera, and K. Y. Eg, "Navigating the landscape of concept-supported XAI: Challenges, innovations, and future directions," *Multimedia Tools and Applications,* pp. 1-51, 2024.

[32] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International Conference on Machine Learning*, 2020: PMLR, pp. 5338-5348.

[33] M. Rigotti, C. Miksovic, I. Giurgiu, T. Gschwind, and P. Scotton, "Attention-based interpretability with concept transformers," in *International Conference on Learning Representations*, 2021.

[34] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viegas, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*, 2018: PMLR, pp. 2668-2677.

[35] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," *Advances in Neural Information Processing Systems,* vol. 32, 2019.

[36] R. Zhang, P. Madumal, T. Miller, K. A. Ehinger, and B. I. Rubinstein, "Invertible concept-based explanations for cnn models with non-negative concept activation vectors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, no. 13, pp. 11682-11690.

[37] J. Crabbé and M. van der Schaar, "Concept Activation Regions: A Generalized Framework For Concept-Based Explanations," *arXiv preprint arXiv:2209.11222,* 2022.

[38] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096,* 2018.

[39] D. Hughes and M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," *arXiv preprint arXiv:1511.08060,* 2015.

[40] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *31st International Conference on Neural Information Processing Systems,* vol. 30, pp. 6629-6640, 2017.

[41] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Structural similarity index (SSIM) revisited: A data-driven approach," *Expert Systems with Applications,* vol. 189, p. 116087, 2022.

[42] O. Keleş, M. A. Yılmaz, A. M. Tekalp, C. Korkmaz, and Z. Doğan, "On the Computation of PSNR for a Set of Images or Video," in *2021 Picture Coding Symposium (PCS)*, 2021: IEEE, pp. 1-5.

[43] Y. Yu, H. Jiang, X. Zhang, and Y. Chen, "Identifying Irregular Potatoes Using Hausdorff Distance and Intersection over Union," *Sensors (Basel),* vol. 22, no. 15, p. 5740, Jul 31 2022, doi: 10.3390/s22155740.